Table of Contents

[Summary: 2](#_Toc19805723)

[Boxplot: 2](#_Toc19805724)

[Feature Selection: 3](#_Toc19805725)

[Model Evaluation: 3](#_Toc19805726)

[Model 1: 3](#_Toc19805727)

[Model 2: 5](#_Toc19805728)

[Comparing Model1 and Model 2: 6](#_Toc19805729)

[Conclusion: 6](#_Toc19805730)

[Summary: 7](#_Toc19805731)

[Boxplot: 7](#_Toc19805732)

[Train and Test Data: 8](#_Toc19805733)

[Model 1: 8](#_Toc19805734)

[Model 2: 10](#_Toc19805735)

[Comparing ROC curves: 11](#_Toc19805736)

[Tabulation: 12](#_Toc19805737)

[Conclusion: 12](#_Toc19805738)

Question No 1 :

Whether the client has subscribed a term deposit or not

Available columns are: age","job","marital", "education", "default" ,"balance", "housing", "loan", "contact", "day", "month", "duration", "campaign", "pdays", "previous", "poutcome", "y"

Target Variable is: “y” i.e. in Categorical

## 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 |

We can see significant difference between the mean and median of some of the 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 |  |

Here in the column Default and y, the categories are not balanced, and as it’s a natural data we can’t

|  |  |
| --- | --- |
| 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 |

## Boxplot:

## Feature Selection:

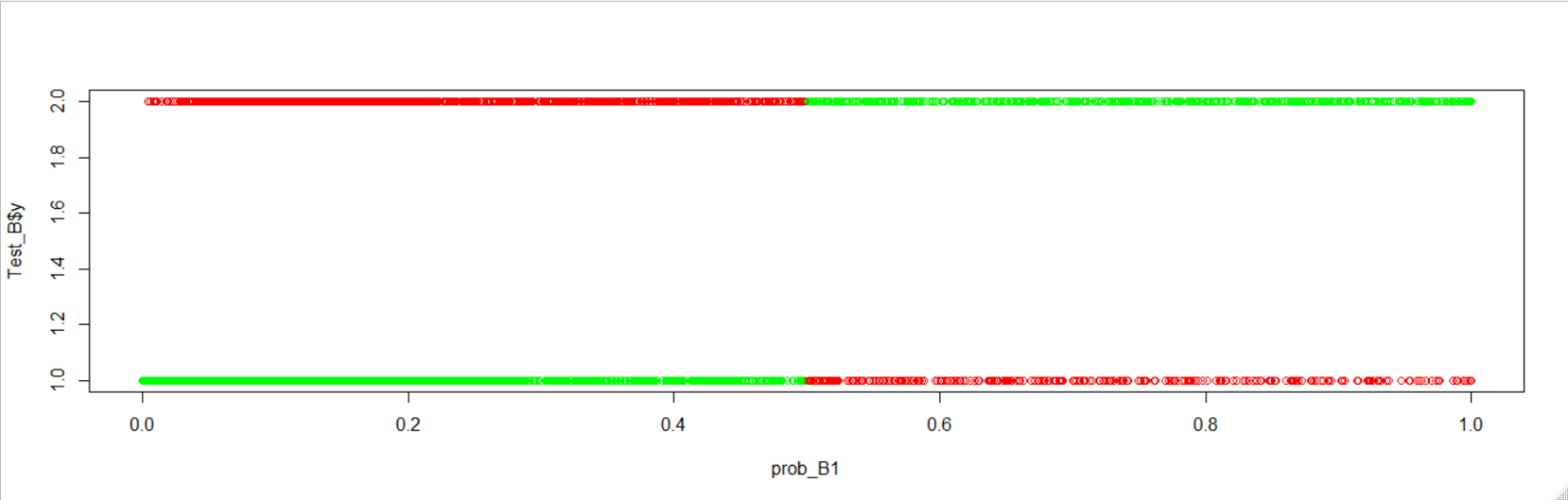
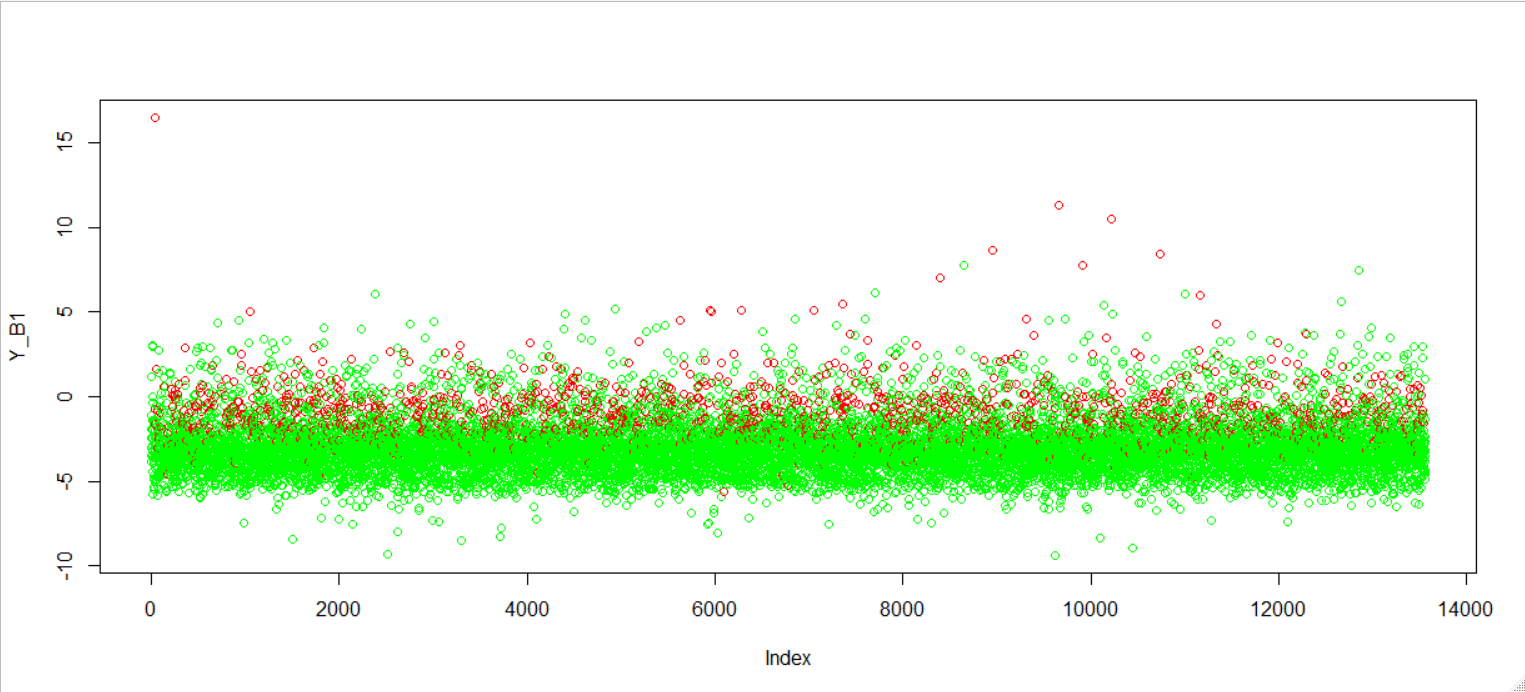
Train data set contains 31648 records and Test data set contains 13563 records

## Model Evaluation:

### Model 1:

In model 1 I have considered all the columns as well as all the Train data records in my model.

model\_B1 <- glm(y~.,data = Train\_B,family = binomial(link = "logit")) Here we get AIC = 15017

Plotting the prediction of the model below, where red means wrong prediction and green points means actual prediction.

Confusion Matrix:

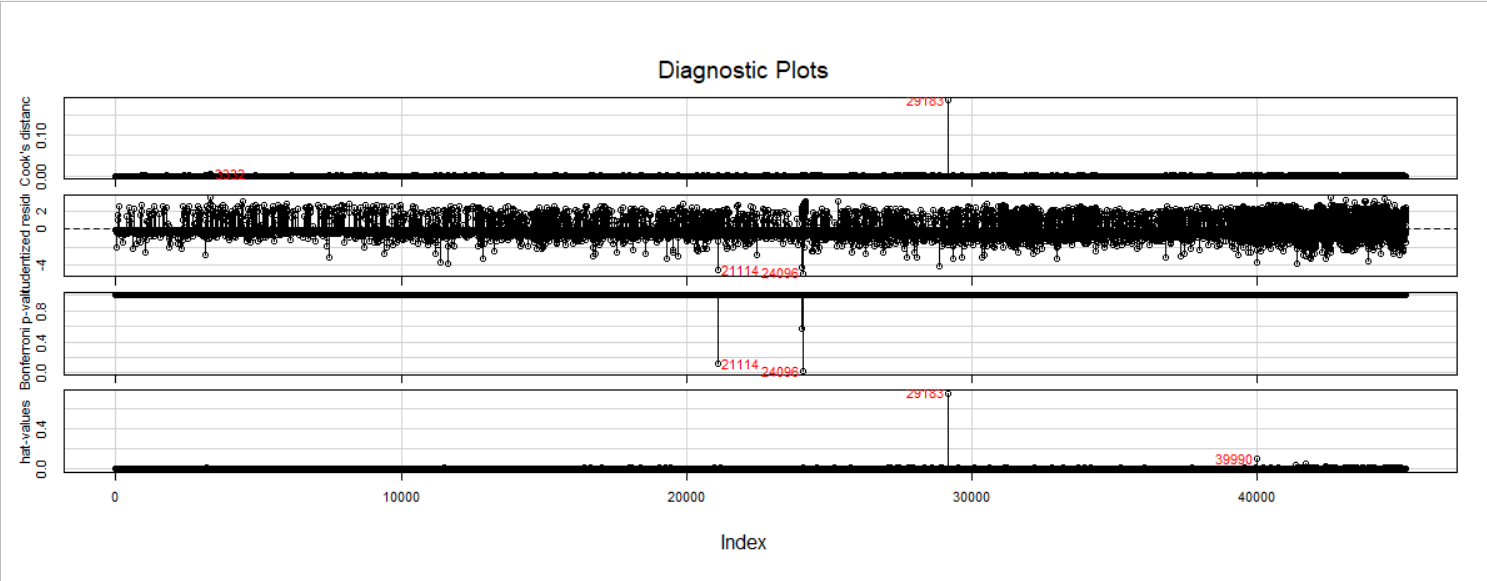
no yes

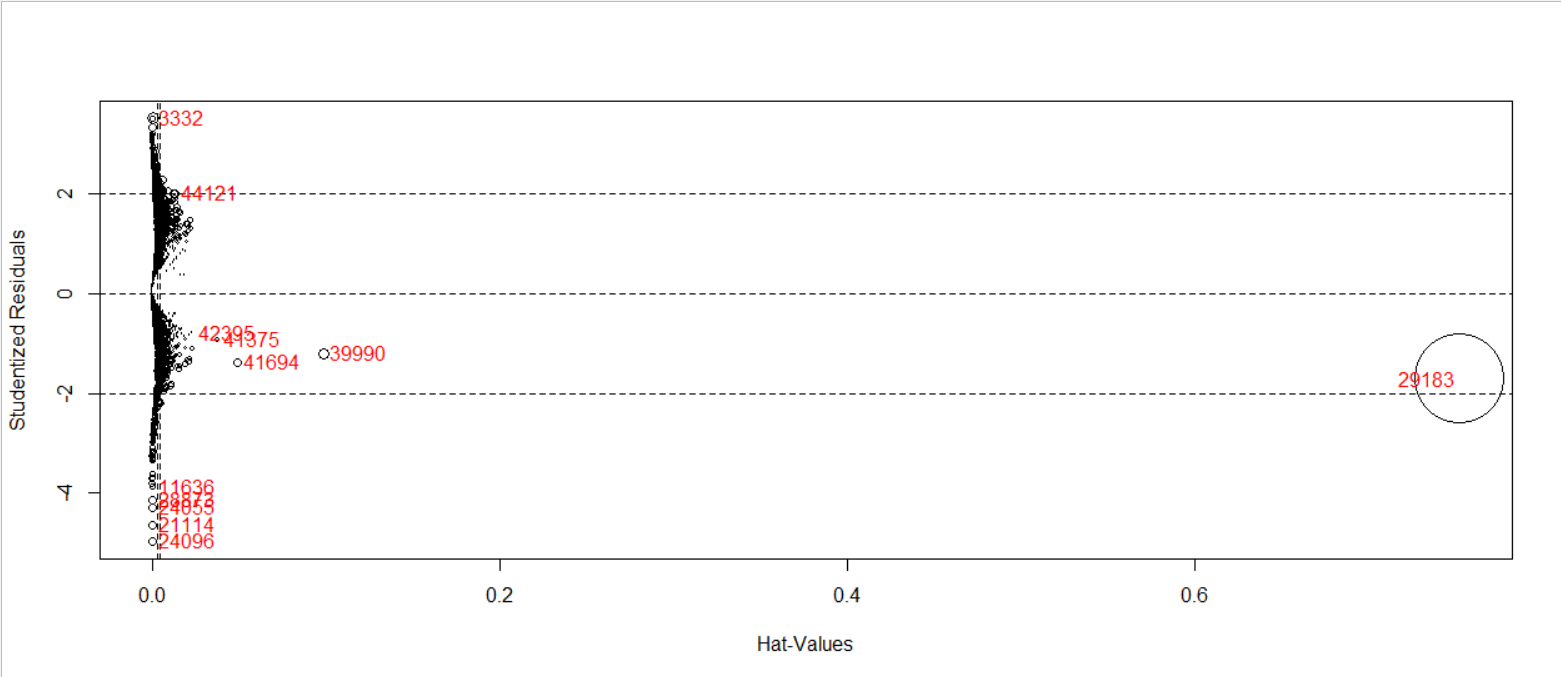
FALSE 11660 1065

TRUE 287 551

Efficiency: 0.900317

Influence Plot:



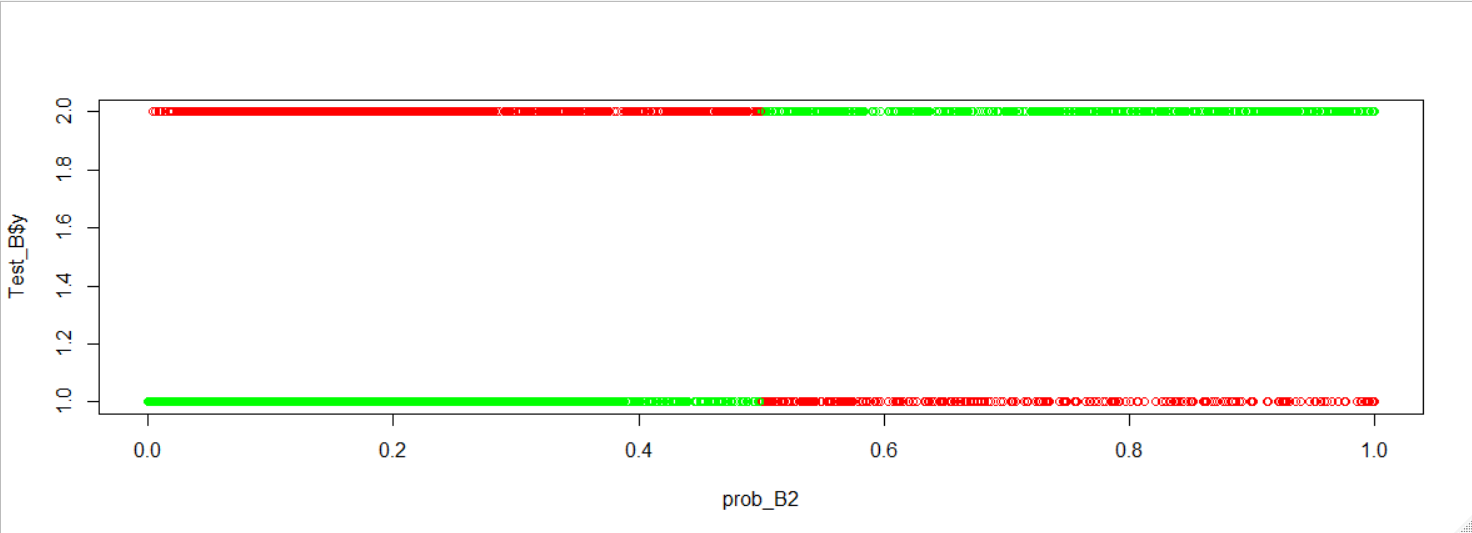
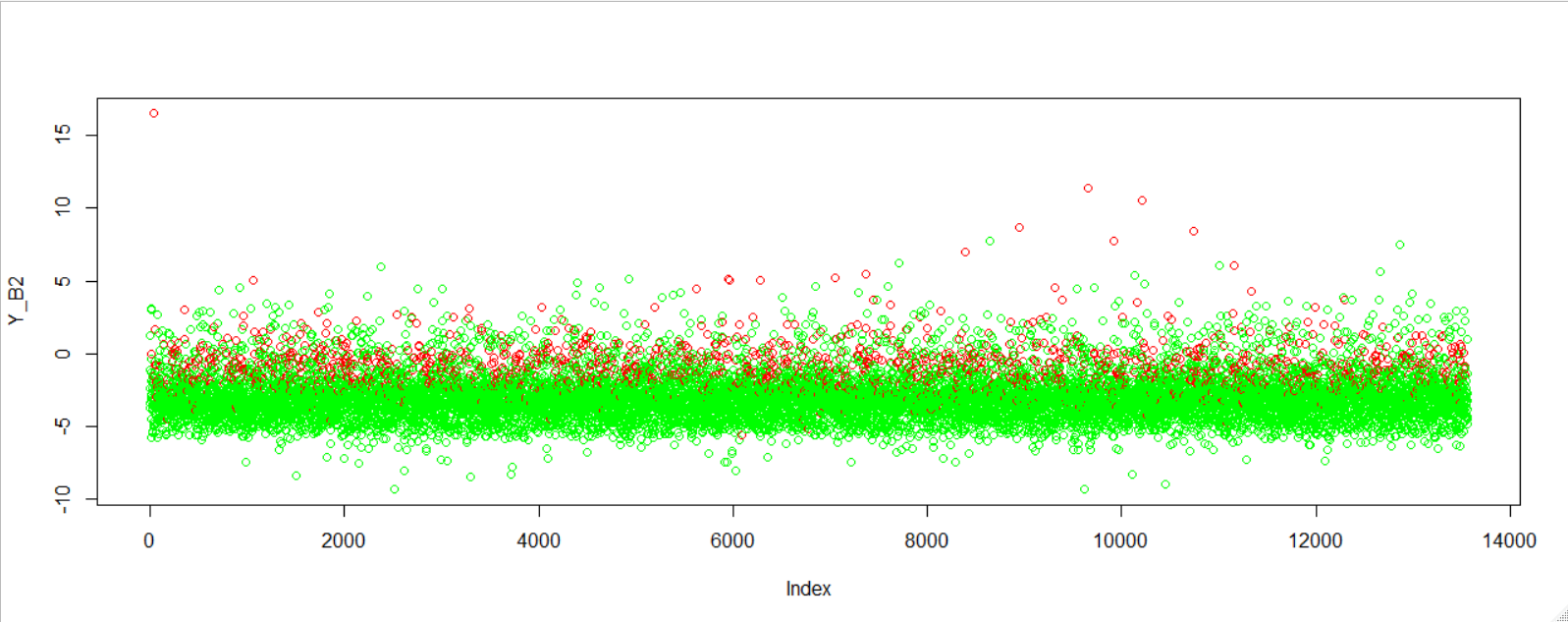


### Model 2:

In model 2 I removed some of the influencing records. And removed some of the insignificant columns.

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

Where I got AIC value as 15010

Confusion Matrix:

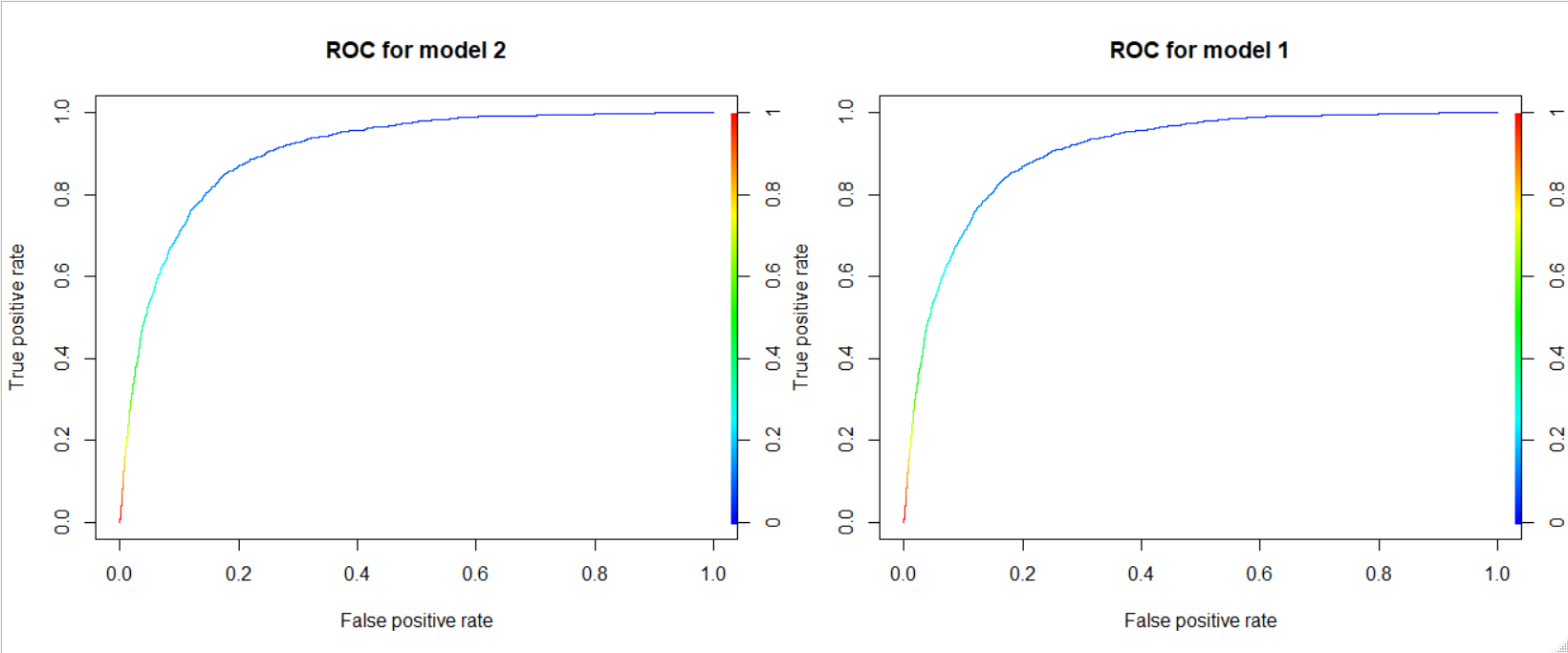
no yes

FALSE 11659 1067

TRUE 288 549

Efficiency: 0.9000958

## Comparing Model1 and Model 2:



|  |  |  |  |
| --- | --- | --- | --- |
| Model No | AIC | Efficiency | F1 Score |
| Model 1 | 15017 | 0.900317 | 0.945201 |
| Model 2 | 15010 | 0.9000958 | 0.9450817 |

## Conclusion:

Here In model 1 and model 2 we can’t see any major differences in our AIC, Efficiency as well as F1Score in both of the models up to 3 decimal point is almost same. As we know in our model 1 we have considered many insignificant variables and in model 2 we have considered only the significant variables for our model building. So I may consider my Model 2 as my final model.

Question 2:

I have a dataset containing family information of married couples, which have around 10 variables & 600+ observations. Independent variables are ~ gender, age, years married, children, religion etc. I have one response variable which is number of extra marital affairs. Now, I want to know what all factor influence the chances of extra marital affair. Since extra marital affair is a binary variable (either a person will have or not), so we can fit logistic regression model here to predict the probability of extra marital affair.

Answer:

Available Columns:

"X", "gender", "age", "yearsmarried", "children", "religiousness", "education", "occupation", "rating", "EMA"

Target Variable is “EMA” Extra Marital Affair, which is a categorical variable with values “yes” and “no”.

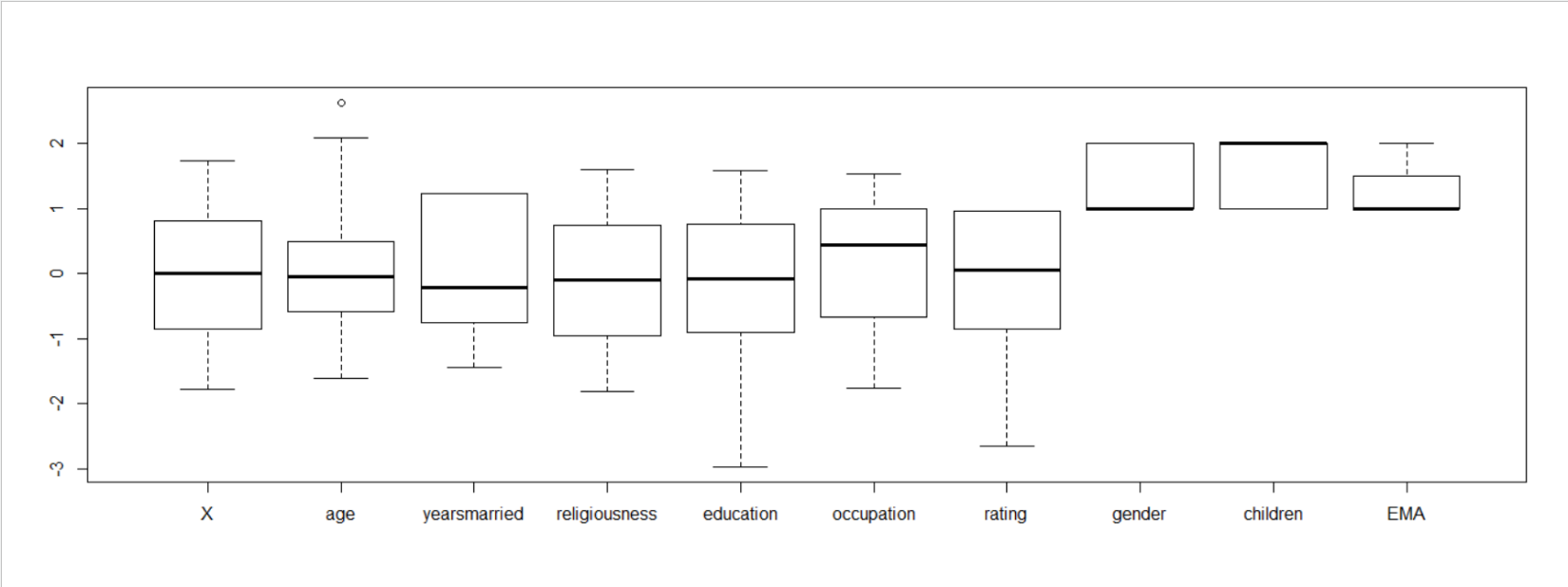
## Summary:

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| X | age | Years married | religiousness | education | occupation | rating |
| Min. : 4.0 | Min. :17.50 | Min. : 0.125 | Min. :1.000 | Min. : 9.00 | Min. :1.0 | Min. :1.000 |
| 1st Qu.: 524.5 | 1st Qu.:27.00 | 1st Qu.: 4.000 | 1st Qu.:2.000 | 1st Qu.:14.00 | 1st Qu.:3.0 | 1st Qu.:3.000 |
| Median : 998.5 | Median :32.00 | Median : 7.000 | Median :3.000 | Median :16.00 | Median :5.0 | Median :4.000 |
| Mean : 993.0 | Mean :32.46 | Mean : 8.147 | Mean :3.119 | Mean :16.17 | Mean :4.2 | Mean :3.933 |
| 3rd Qu.:1447.0 | 3rd Qu.:37.00 | 3rd Qu.:15.000 | 3rd Qu.:4.000 | 3rd Qu.:18.00 | 3rd Qu.:6.0 | 3rd Qu.:5.000 |
| Max. :1960.0 | Max. :57.00 | Max. :15.000 | Max. :5.000 | Max. :20.00 | Max. :7.0 | Max. :5.000 |

|  |  |  |
| --- | --- | --- |
| EMA | gender | children |
| no :447 | female:313 | no :170 |
| yes:149 | male :283 | yes:426 |

From the summary we can say, possible There are negligible difference between the median and mean, may be data contains very a smaller number of outliers.

## Boxplot:

From boxplot we can say that age variable contains outlier.

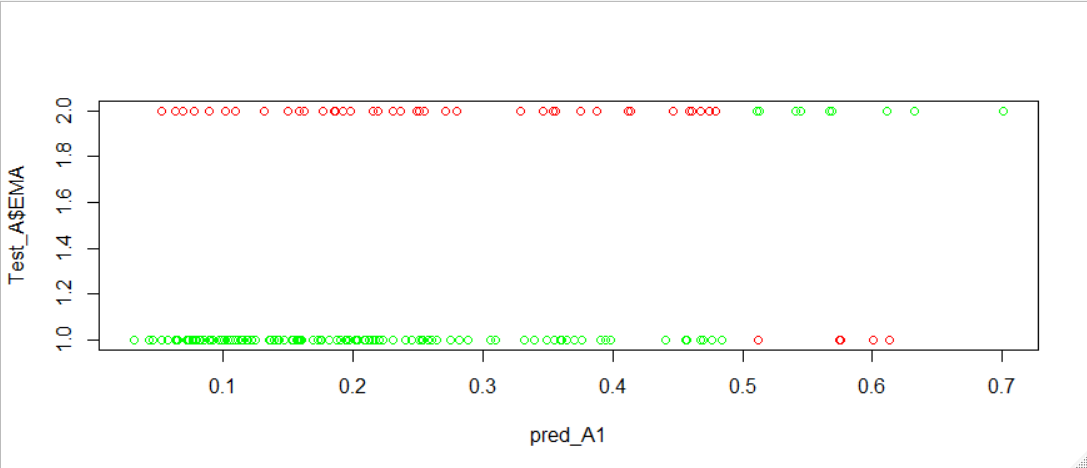
## Train and Test Data:

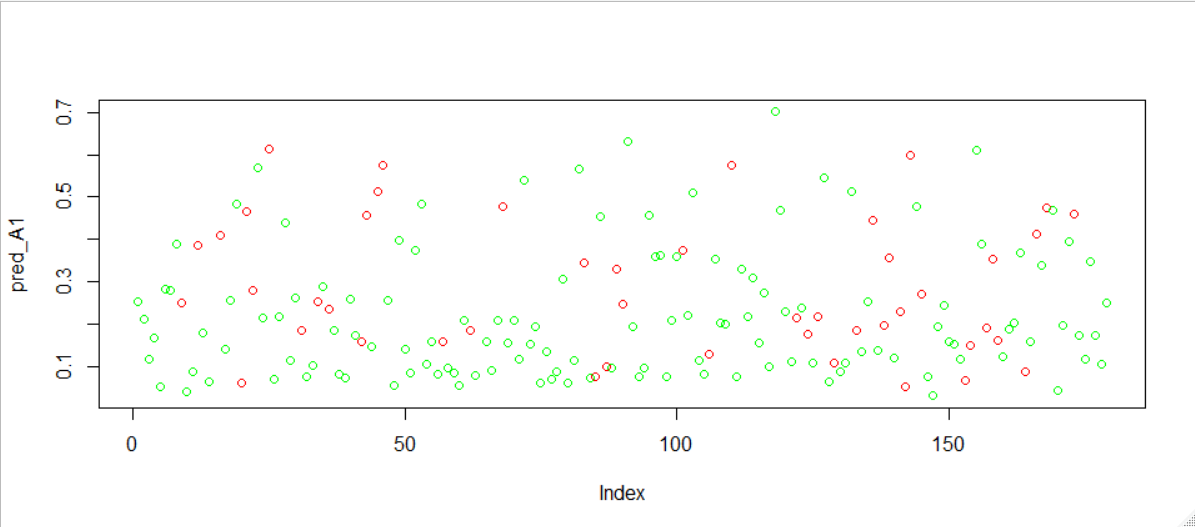
Train Data contains 419 records and my test data contains 179 records in my data.

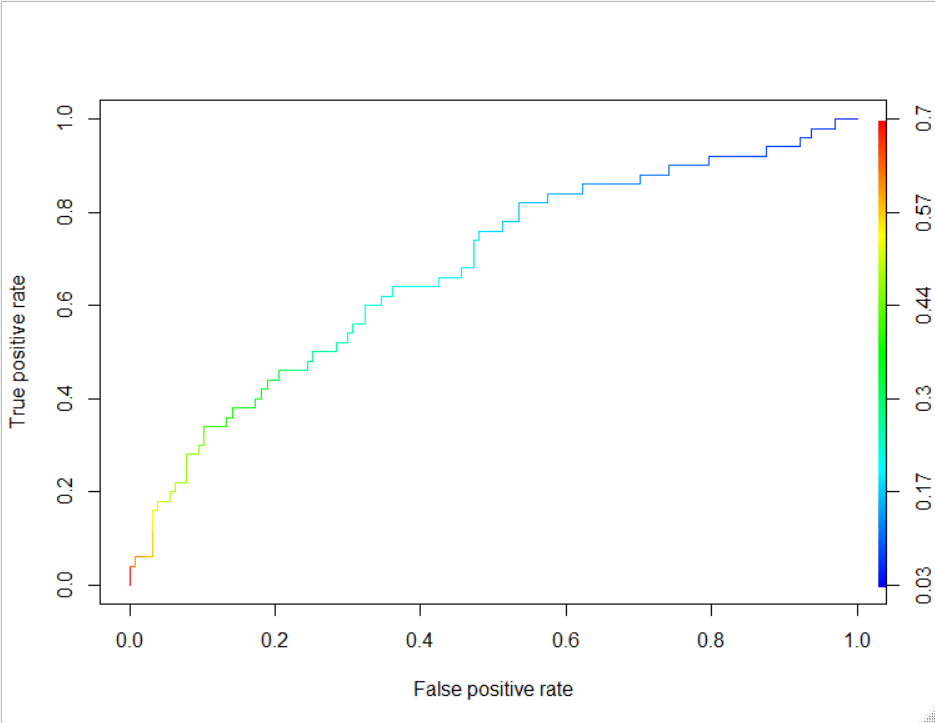
Model Building:

### Model 1:

In my model 1 i have considered all my variables and records from the Train data, where I get my AIC value as 429.34, and efficiency is 0.740113.

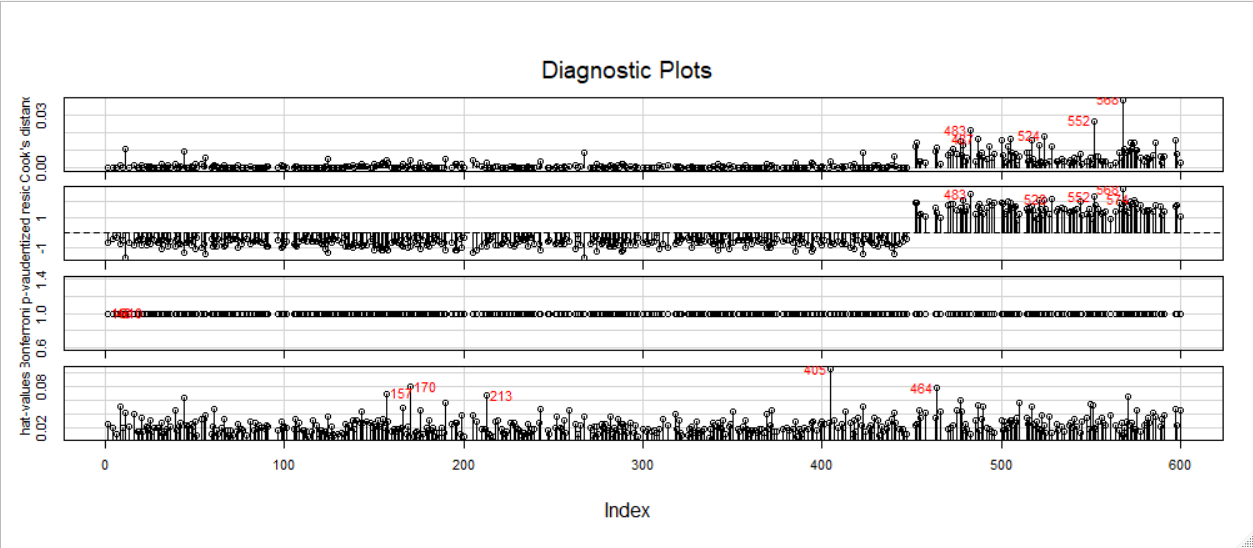


The red marks are wrong prediction and the green marks are correct prediction using my model 1.

ROC Curve:

Here we can see that the area under the curve is very less.

Influence Plot:

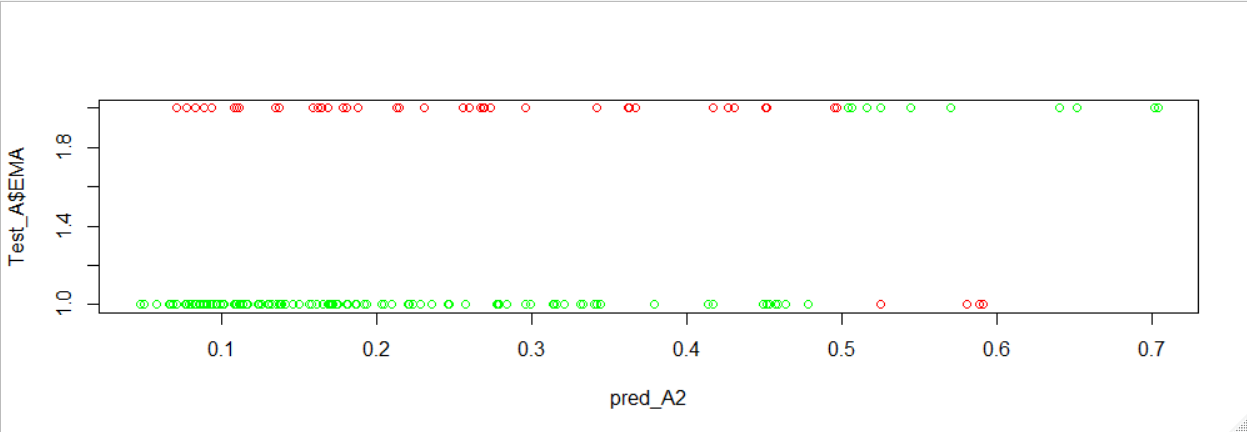
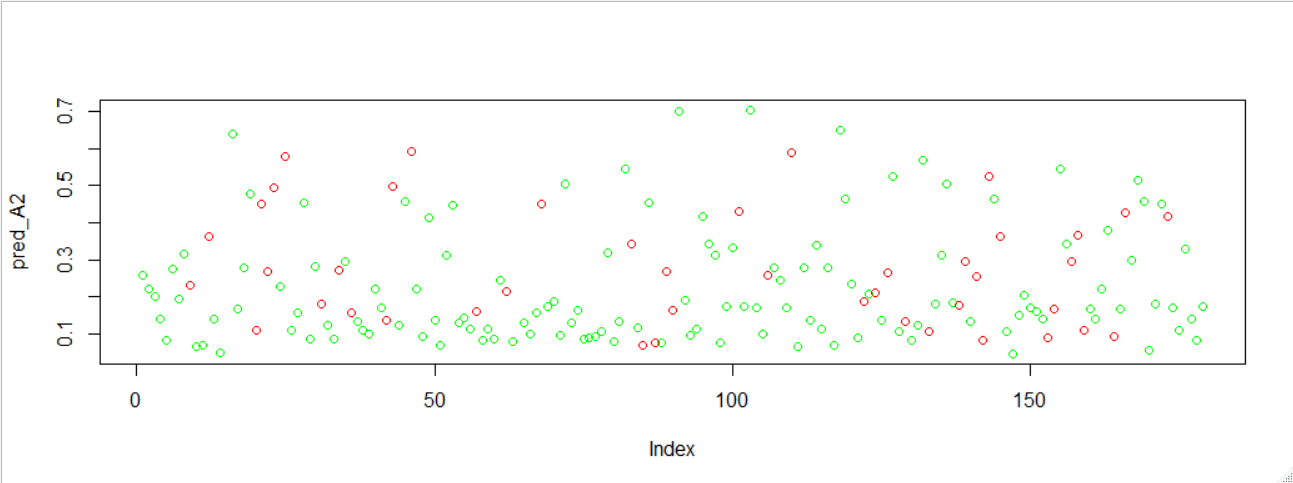


We can see some Influencing value in our model, so we may remove the Influencing Records for my next model.

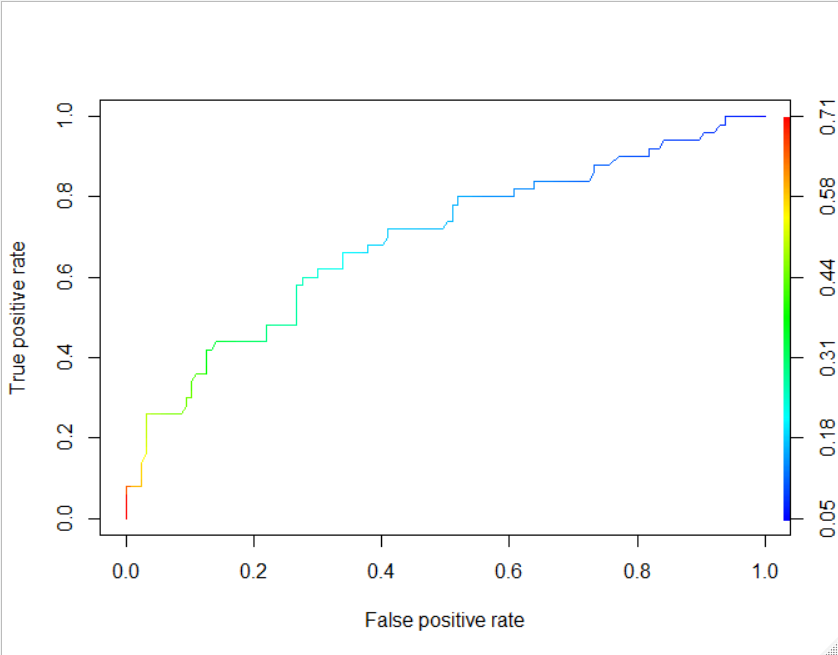
### Model 2:

I modeled my second model with only the significant columns and removing the influence index. As I have tested the model with all the columns and removing the influence index once, and come up with conclusion that, even if I remove the influence index, the variables are still insignificant, so in my model 2 I removed both influencing records as well as the insignificant columns.

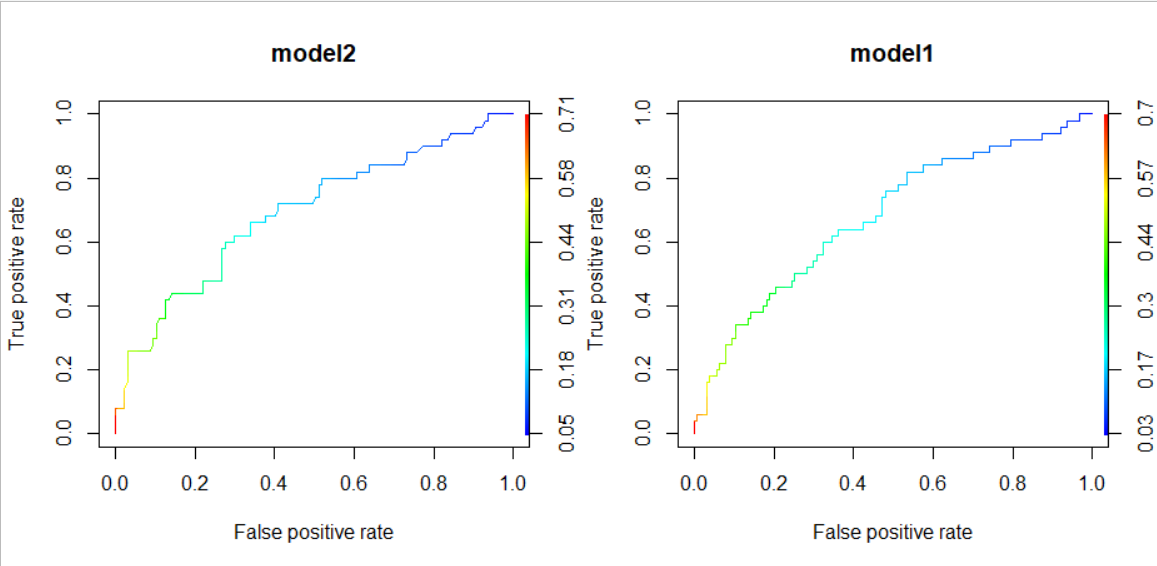
Here in model 2 I come up with AIC as 421.8 which is less than our previous model, and efficiency as 0.7570621 i.e. little bit increased.



ROC Curve:



## Comparing ROC curves:



Looking at the model 2 we can say that the area under the ROC curve is increased in model 2 as compare to model 1.

## Tabulation:

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

## Conclusion:

From the ROC curve and the above tabulation, I come up with conclusion that, in our model 2 we are getting less AIC, Higher Efficiency, and Higher F1 Score in our model 2. So I may prefer to go for my model2.