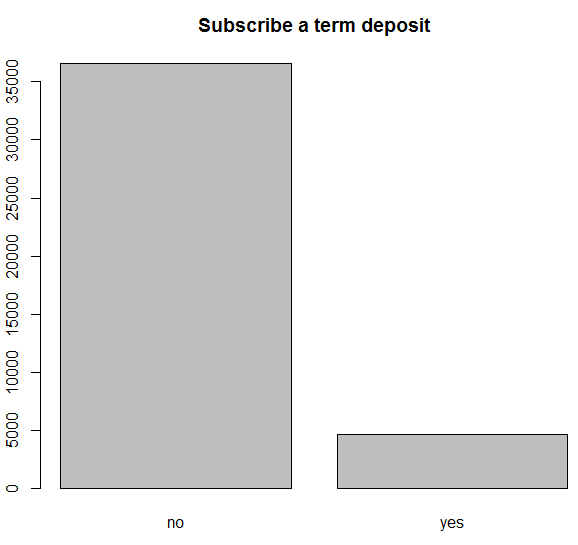
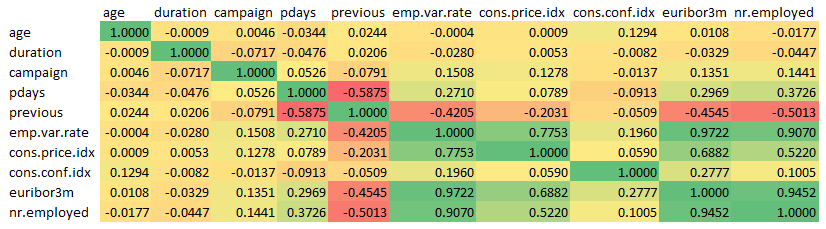
**Bank marketing – Analysis**

**The spread of the dependent variable Y**



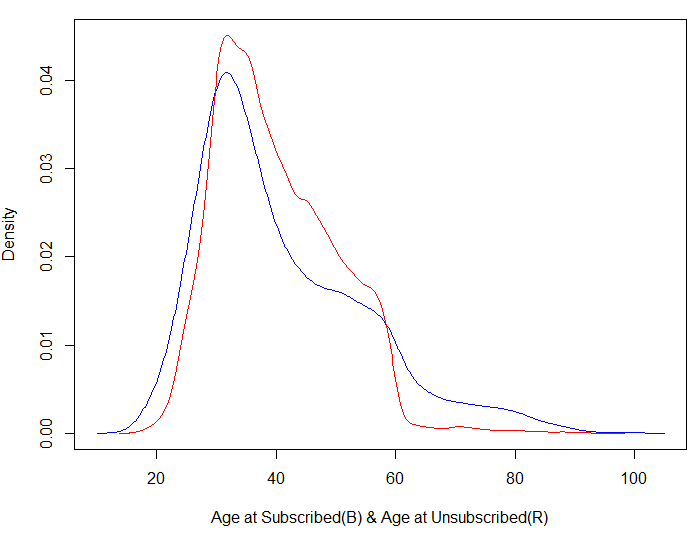
**Correlation matrix**



**Points:**

* ‘Previous’ feature has a Negative Correlation with Pdays, emp.var.rate, euribor3m & nr.employed
* Emp.var.rate , cons.price.idx, Euribor3m and nr.employed are correlated with each other

Here is a density plot that shows how the Age feature is split across the loan subscribed and loan unsubscribed

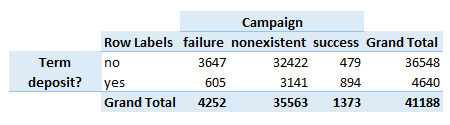


This density plot shows the Age of people who has subscribed the loan (In Blue) and those who have not subscribed the loan (In Red).

**Points:**

* Those who have not subscribed are mostly between the Age of 20 to 60
* And those who have subscribed are teens till very old people, which means in almost every age criteria there are people who has subscribed the loan.

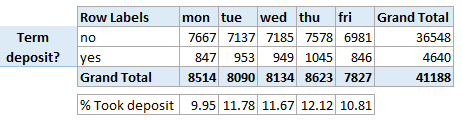
**Outcome of previous marketing campaign**



**Points:**

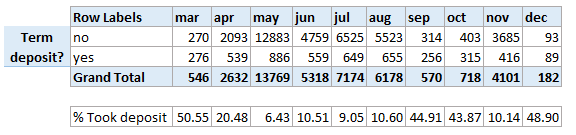
* In the above table, when the previous campaign was successful, 894 people out of 1373 (64%) have taken the term deposit.
* Hence marketing campaign benefits the Bank.

**Day of the week vs deposit taken**



It looks like only 10% of deposits were taken every day of the week.

**Month-wise count of term deposits**

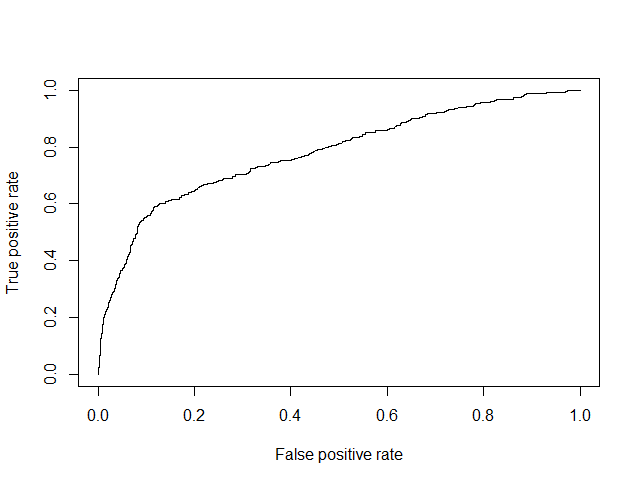


**Points:**

* During March, Sept, Oct and Dec there were around 50% of term deposits taken
* Whereas during other months, only 10% of them has taken deposits

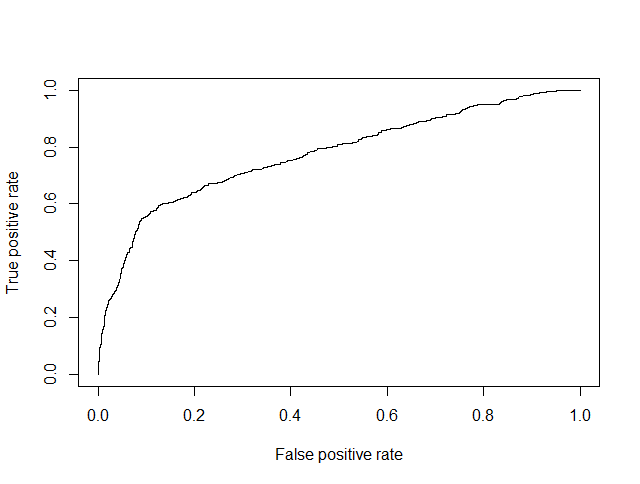
ROC Curve:

This is the ROC curve of the logistic model including the multi-collinear features



The ROC Curve will go to the top left corner for a good model. This is a good model even though we have around 20 independent variables.

Now , on omitting the multi-collinear variables, we created a model, predicted it and below is the ROC curve



The ROC curve almost looks the same as that of the initial curve. Hence there is not much change.

**Confusion matrix**

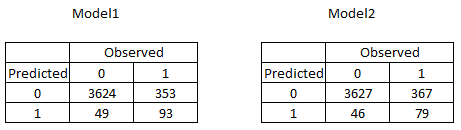
1 – Customer will subscribe loan

0 – Customer will not subscribe loan

* Logistic models:

Model 1 – with multi-collinear variables

Model 2 – without multi-collinear variables

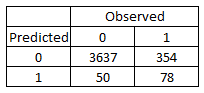


The trickiest part of confusion matrix is False Negative – Predicting that I will take the subsidy and I actually will not. This is very important of the confusion matrix – the lesser the value the better.

Even though the True Positive is more in Model1, False Negative is less in Model2 which is better.

Accuracy : Model 1 – 90.24% , Model 2 – 89.97%

* Support Vector Machine:

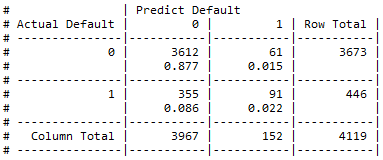


Comparing this with Logistic Model 2, we have 50 here whereas logistic model has 46. Hence we go with Logistic model

Accuracy : 90.19%

* Decision Trees:

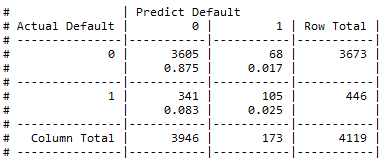
Ordinary Decision tree:



Accuracy : 89.90%

Decision Tree accuracy is not better compared to Logistic model and SVM

Decision Tree Boost model:



Accuracy : 90.07%

Comparing confusion matrix values of all the models, we can see SVM model performs better. So we finally use the predictions of SVM model.