Project _2_Report

This project relates to the Direct Marketing exercise done by a Portuguese banking institution.

The marketing campaigns were based on phone calls. Often, more than one contact to the same client was required,

in order to access if the product (bank term deposit) would be (or not) subscribed.

The classification goal is to predict if the client will subscribe a term deposit (variable y).

Output variable (desired target):

Has the client subscribed a term deposit? (Binary: **"yes","no"**).

The output variable **yes or no** is classified in the project based on **probability values.**

Given by the **logistic regression model.**

(Set Cutoff probability = 0.5). (yes=1, no=0).

If probability > 0.5 we classify it as **yes** that means the client subscribed a term deposit. Otherwise not .Accordingly we predict the future (yes/no) using the model given in the project.

We start the analysis by bringing the data in r and cleaning it and removing NA values

But we don’t find any NA values in data.

Next, we run the logistic regression on the data with 16 independent variable and dependent variable y.

We ran the model summary in which we find that previous and pdays are not significant

Also we had a look at other summary like null deviance, residual deviance, AIC and we had a look at predicted values given by the model.

We also plotted the predicted (fitted) values.

Secondly, after removing all insignificant variable and checking significance of factor Variable using Wald's test. We develop another model model 2, and checked its Summary , vif , deviance, AIC etc

Further we divided the given data in to training data and test data , we ran the model 2

model2<-

glm (y~poutcome+campaign+duration+month+day+contact+loan+housing+balance +default+ education+ marital+ job, data=bank, family=binomial (logit))

on training data and checked the model summary found some useful information like null deviance, Residual deviance , found that residual deviance is low as compared to null deviance which is good , also we check the AIC of the model , ran the Anova test, Also there is no mullticollinearity in the model.

After that we have predicted the values for the model on train set .

Later, we ran the model 2 on test set and gone through all procedures which we did for the train set.

Finally we check the performance of model 2 on both train set, using performance Measures like, Confusion Matrix, ROC plot, Area under Roc curve, Cross-validation Accuracy .

We found the following numbers

We have better performance of the confusion matrix at the cutoff of 0.5

Confusion matrix:-

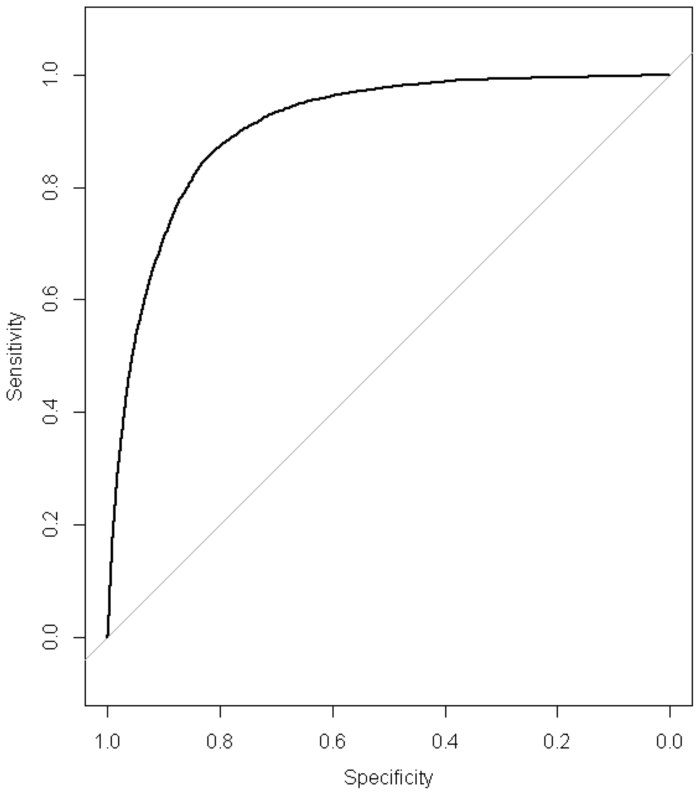
```
Threshhold =  0.5
```

```
Confusion Matrix :
```

```
class1     no   yes
     0 31215  2715
     1   774  1464
```

```
% correct = 90.4
--------------------------
False Positive Rate = 0.08
--------------------------
False Negative Rate = 0.08
--------------------------
```

ROC Curve :-



The curve above seems to be tented towards left hand corner which shows that

The model has good classification ability.

```
Area under the curve: 0.9079
```

The Performance of model 2 on testset

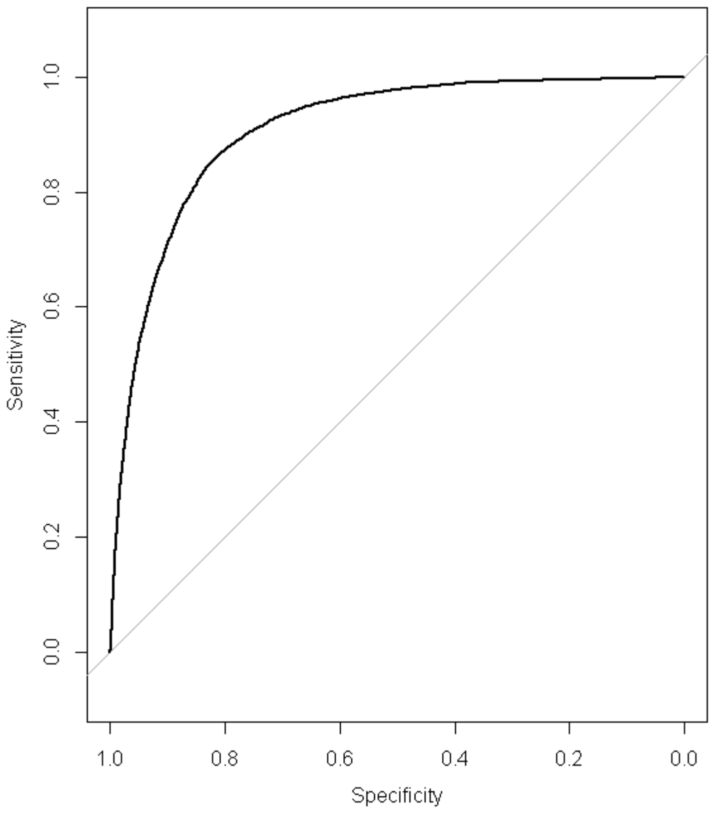
```
Threshhold =  0.5

Confusion Matrix :

class1     no   yes
     0 31215  2715
     1   774  1464

% correct = 90.4
-----------------------------
False Positive Rate = 0.08
-----------------------------
False Negative Rate = 0.08
-----------------------------
```

The Roc Curve:-



```
Area under the curve: 0.9079
```

Overall we found that the performance of model 2 on test set and train set is

Which suggest that we have pretty competitive model .

Finally we also got the cross validation accuracy of the model

```
Fold:  4 3 2 10 8 9 6 7 1 5
Internal estimate of accuracy = 0.902
Cross-validation estimate of accuracy = 0.902
```

This completes the Project.