# Logistic Regression (Module -9)

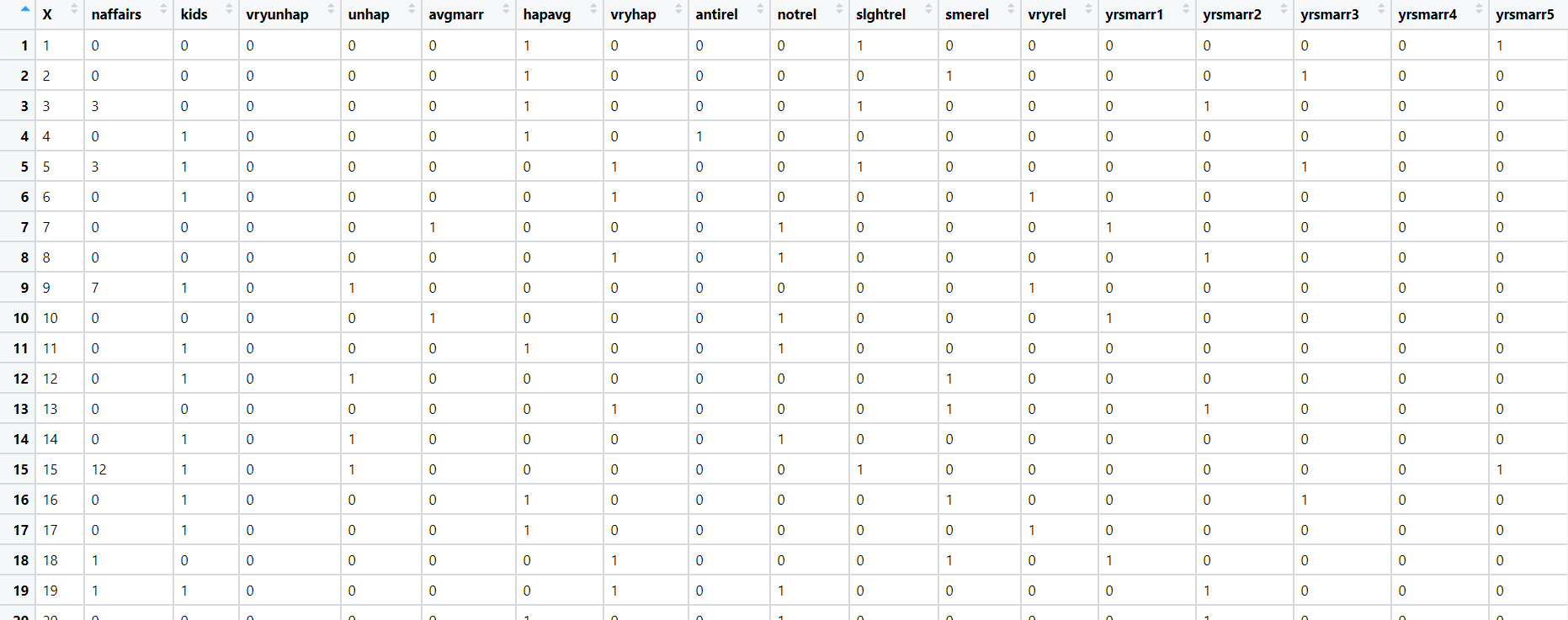
1.) I have a dataset containing family information of married couples, which have around 10 variables & 600+ observations.

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.

**install.packages('AER') data(Affairs,package="AER")**



**Ans:**

**Analysis:**

1. As the output variable is categorical, we need to got with multiple logit regression
2. The Output Variable is already a factor but with different categorical levels
3. Now to apply this on Binomial we need to binarize the variables

**Null values:**

There are no Null values to impute so we can go further with model building

**Building the Model**: Linear Model

1.model <- lm(affairs ~ ., data = Affairs) #609 is the residual deviance

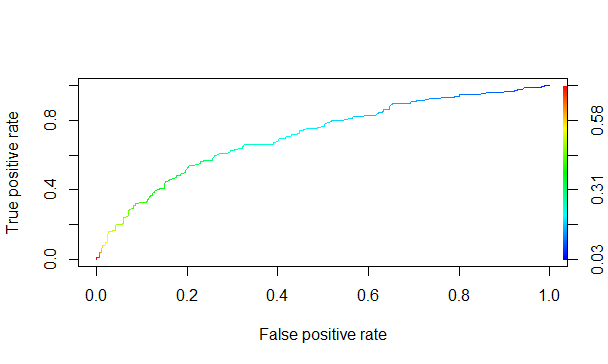
summary(model) #education, occupation, childreneyes, gendermale

2.Binomial Model: model1 <- glm(affairs ~ .-education-occupation, data = Affairs, family = "binomial")

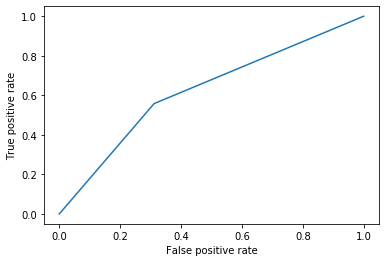
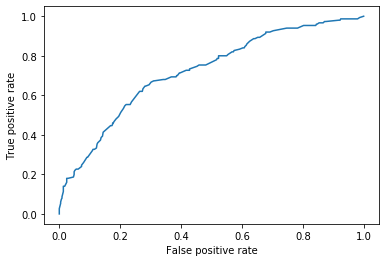
summary(model1) # 610 is the residual deviance

Calculate the Optimal cutoff and Plot the ROC.

Opt\_Cutoff # 0.517 is the optimal cutoff and the Accuracy is 0.768

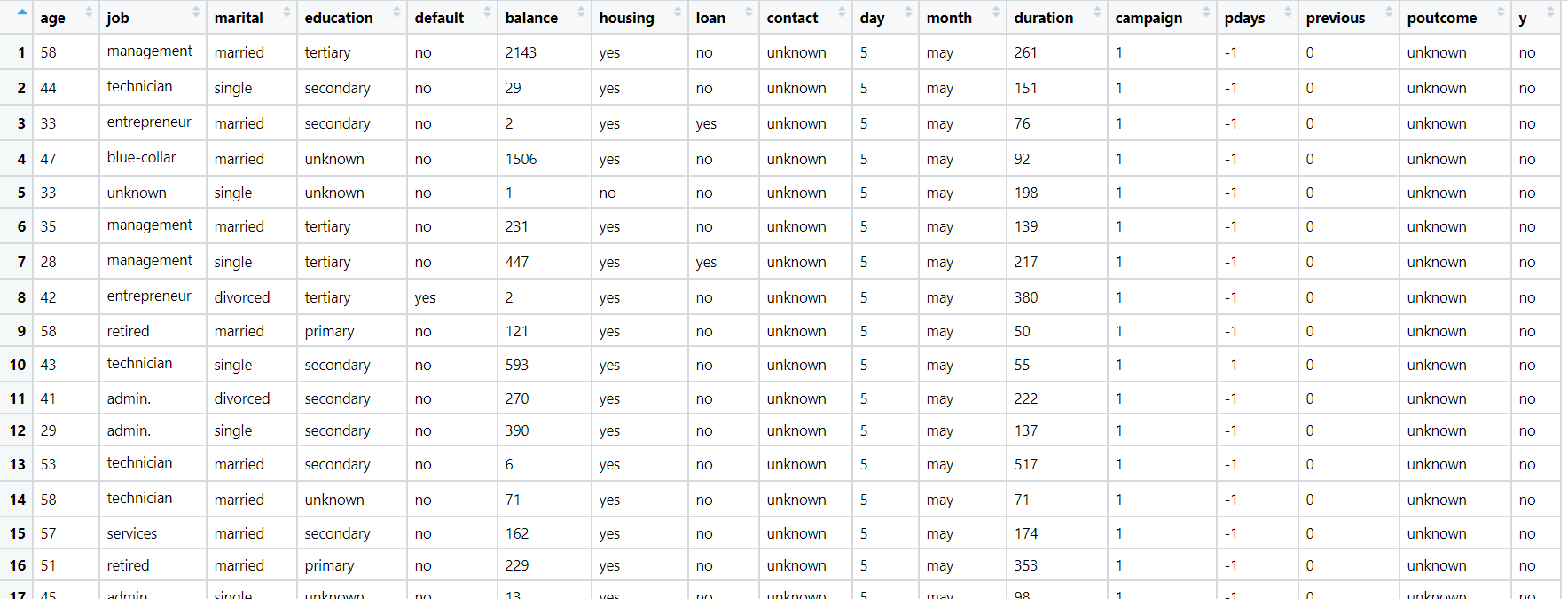


**Roc Curve:**

****

2.) Output variable -> y

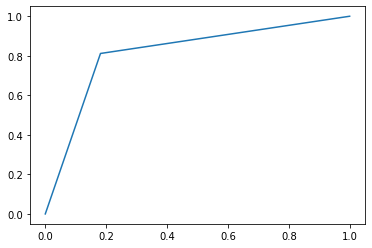
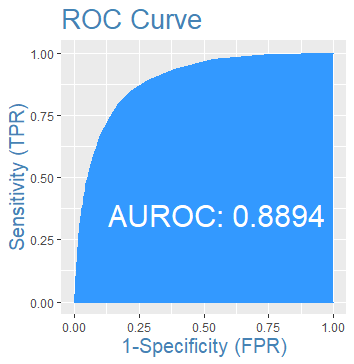
y -> Whether the client has subscribed a term deposit or not Binomial ("yes" or "no")



**Ans:**

**Analysis:**

1. Model Building is done after solving the problem of collinearity and checked with VIF and alias functions
2. Now use the final model for the Analysis to plot the ROC and to calculate the Accuracy value.



model2 <- glm(y ~ .-poutunknown-con\_unknown-single-jounknown-joretired, data = bankdata, family = "binomial")

vif(model2) # Now the coercion is removed with ignoring the joretired value

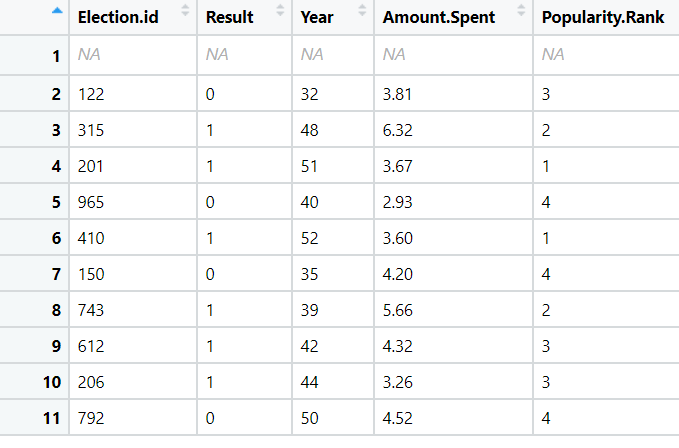
1. Optimal cutoff is 0.5199

1. Accuracy is 0.90

3.) Suppose we are interested in the factors that influence whether a political candidate wins an election.

The outcome (response) variable is binary (0/1); win or lose.

The predictor variables of interest are the amount of money spent on the campaign, the amount of time spent campaigning negatively and whether or not the candidate is an incumbent.



**Analysis:** First Row consists of NA is removed using na.omit()

**Final Model:**

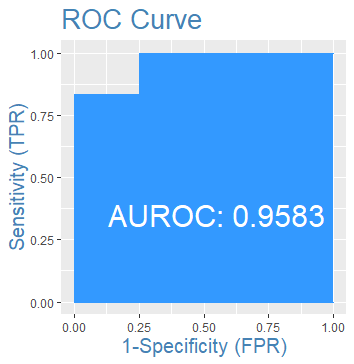
Model1 <- glm(Result ~.- Election.id - Year , data = Election1)

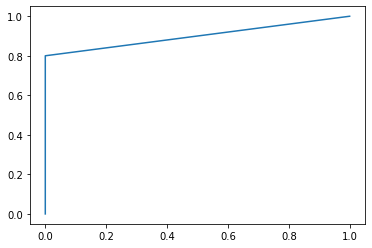
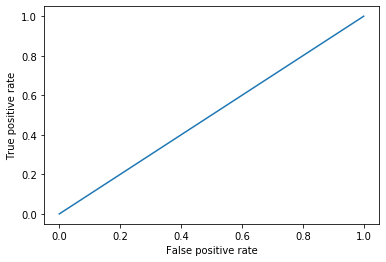
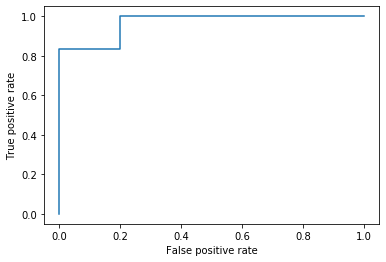
summary (Model1)

Vif is also not greater than 10, there is no collinearity but all the values are insignificant

We removed Year and Election Id then the Residual deviance and significance of other variables are increased.

**ROC:**



****

**Accuracy is 0.9 .**

# Hints:

1. Business Problem
   1. Objective
   2. Constraints (if any)
2. Data Pre-processing

2.1 Data cleaning, Feature Engineering, EDA etc.

1. Model Building
   1. Partition the dataset
   2. Model(s) - Reasons to choose any algorithm
   3. Model(s) Improvement steps
   4. Model Evaluation
   5. Python and R codes
2. Deployment

4.1 Deploy solutions using R shiny and Python Flask.

1. Result Share the benefits/impact of the solution - how or in what way the business (client) gets benefit from the solution provided.

**Note:**

1. For each assignment the solution should be submitted in the format
2. Research and Perform all possible steps for improving the model(s) accuracy.

Ex: Transformations, Feature Engineering, Hyper Parameter tuning, Outlier treatment, etc.

1. All the codes (executable programs) are running without errors
2. Documentation of the module should be submitted along with R & Python codes, elaborating on every steps mentioned here.