**Documentation of Logistic Regression Assignment**

Question 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")

**1. Business Problem**: Predict the possiblilty of a person having an affair

* 1. Objective: Develop a logistic regression Model to predict the possibility of extra marital affair of a person.
  2. Constrains : Nil

1. Data Pre- Processing: The file affairs contains the family information of married couple, which have 9 variables and 600+ observations.

**EDA**: Independent variables are ~ gender, age, years married, children, religion etc. We have one response variable which is number of extra marital affairs. Now, We 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.

1. Model Building

Logistic Regression Model

glm function to use sigmoid curve

model <- glm(ynaffairs ~ factor(gender) + age+ yearsmarried+ factor(children) + religiousness+

education+occupation+rating, data = affairs1,family = "binomial")

Confusion matrix table

prob <- predict(model,affairs1,type="response")

summary(model)

Model Accuracy = 76.91 %

**Question 2**

Output variable -> y

y -> Whether the client has subscribed a term deposit or not

Binomial ("yes" or "no")

**Business Problem**​: Predict if the client has subscribed to a term deposit or not.

**DATA Pre processing:**

Check for Duplicate Rows

Check for Missing Data

How Many Rows Are Completely Missing Values In All Columns

Assigning ‘yes’ to binary

**Exploratory Data Analysis**

Univariate Analysis

Age Distribution

The bulk of clients are between the ages of 33 (1st Quartile) and 48 (3rd Quartile) with mean lying on 41 visualized on the histogram with red vertical line.

Boxplot of age describes essentially the same statistics but we can see outliers above the age of 65.

**Model Building**

Scatterplot of Duration by Age  
Scatterplot of Duration by Campaign

Scatterplot Matrix

Split the Training and Testing data as 70 % and 30%

Plotting distribution

Model - Build Model by fitting Logistic Regression

Predict the Logistic Regression model on testing data and evaluate by finding the accuracy and calculating/plotting ROC curve.

**Question 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.

**Business Problem**: To predict if a candidate wins or loses in the elections

**About the data**: The Election file data contains the result whether a political candidate wins in election or not this file has 5 variables and 10 observations.

Here response variable is results it contains win or loss information . The goal is to try to find out the factor which are influenced the results in the given variables that are 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.

**Exploratory Data Analysis**

* Given dataset contains 5 N/A values
* 5 N/A values found in first record only, hence let’s remove first record from the dataset.

#Logistic Model Building

Election ID is not significant for election result hence, preparing model without Election ID

**AIC = 8**

**Confusion Matrix**

True Positive =6 and True Negative =4

Accuracy = TP+TN/TP+TN+FP+FN =(6+4)/( 6+4+0+0)=1

As it is a small data set of 10 records it has given 100% accuracy.

**Conclusion:**

More area under the ROC Curve is better

This model is predicting all the values correct as the dataset is extremely small with only 10 observations