SUMMARY

CLEANING DATA:

Reading in the data set provided and checking for duplicate value. Dealing with missing values and dropping column with over 40% missing values. Then changing missing values percentages for each column into a data frame by choosing columns with missing values between 0 and 15% and we will be dropping the missing values from those columns that have missing values less than 15%. Tags column has 36% missing values so it wouldn't be ideal to impute that many values with the mode or to drop them altogether. Percentage of rows left after data cleaning process is 98 percent. So we still have a lot of the data retained for EDA and model building

EDA:

A quick EDA was done to check the condition of our data. It was found that a lot of elements in the categorical variables were irrelevant. The numeric values were good and only few outliers were found. We decided to cap outliers in column Total Visits to 125 and outliers in column 'Page Views Per Visit' to 25 as we don't want two or three datapoints to skew our analysis. The following are the notable insights for categorical variables:

1. Google and Direct Traffic are the biggest lead sources
2. Overwhelming majority of the leads are Indians
3. More leads have prior experience in Finance, HR and Marketing management
4. A good chunk of the leads are having current status of 'Will revert after reading the Email' or 'Ringing'
5. Majority of the leads are looking for a course for better career prospects.
6. Majority of the leads wants a free copy of 'Mastering the Interview'
7. Majority of the leads are currently unemployed
8. The last notable activities of students/leads are 'Modified', 'Email opened' and 'SMS sent'

TRAIN-TEST SPLIT:

The split was done at 75% and 25% for train and test data respectively.

MODEL PERFORMANCE:

Firstly, RFE was done to attain the 30 relevant variables. Later the rest of the variables were removed manually depending on the VIF values and p-value where it ended up with 17 variables.

Model Evaluation: A confusion matrix was made. Below are the predictions from Confusion Matrix:

- 4032 of them are to be true negatives for train and 1370 of them are to be true negatives for test
- 166 of them are to be false positives for train and 71 of them are to be false positives for test
- 300 of them are to be false negatives for train and 116 of them are to be false negatives for test
- 2307 of them are to be true positives for train and 712 of them are to be true positives for test

Later on the optimum cut off value (using ROC curve) was used to find the accuracy, sensitivity and specificity which came to be as below:

- Overall Accuracy
    - Train : 93.15%
    - Test : 91.75%
- Sensitivity
    - Train : 88.49%
    - Test : 85.99%
- Specificity
    - Train : 96.04%
    - Test : 95.07%

We are getting a precision of around 90.93% or 91% on the test set.