SUMMARY

This analysis is done for X Education and to find ways to get more professionals to join their courses. The basic data provided gave us a lot of information about how the potential customers visit the site, the time they spend there, how they reached the site and the conversion rate.

Basic steps:

- Import Data and all libraries, inspect the data frame
- This helps to give a good idea of the data frames.

The following are the steps used:

1. **Cleaning data:**
   The data was partially clean except for a few null values and the option select had to be replaced with a null value since it did not give us much information. Dropping the columns having more than 40% missing values, impute the missing value with mean & mode. Grouping the response which is having very less count. Few of the null values were changed to 'NA' so as to not lose much data. Although they were later removed while making dummies. Since there were many from India and few from outside, the elements were changed to 'India', 'NA' and 'Outside India'.

2. **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 seem good and no outliers were found. There are outliers in the numeric variables, Capping all numeric variables of these outliers will help in the analysis

3. **Dummy Variables:**
   The dummy variables were created concatenated (joined) the results to the master data frame.

4. **Train-Test split:**
   The split was done at 70% and 30% for train and test data respectively. Perform scaling Divide the data into X and y.

5. **Model Building:**
   Firstly, RFE was done to attain the top 25 relevant variables. Later the rest of the variables were removed manually depending on the VIF values and p-value (The variables with VIF < 5 and p-value < 0.05 were kept). Predicting the probabilities on the train set.

6. **Model Evaluation**:
   A confusion matrix was made. Check the overall accuracy found 81% Later on the optimum cut off value 0.5 (using ROC curve) was used to find the accuracy, The area under ROC curve is 0.88 which is a very good value.

7. **Prediction:**
   Prediction was done on the test data frame and with an optimum With the current cut off as 0.40 we have accuracy: 76.20%, sensitivity: 55.52% and specificity of around 90%.
8. **Precision - Recall**:
   With the current cut off as 0.40 we have Precision around 80.09% and Recall around 66.90%
9. **Precision and recall tradeoff**:
   With the current cut off as 0.42 we have Precision around 73.26% and Recall around 78.02%

Conclusion & Recommendation

It was found that the variables that mattered the most in the potential buyers are (In descending order) :

1. The total time spend on the Website.
2. What is your current occupation housewife
3. When the lead source was:
   a. Google
   b. Direct traffic
   c. Organic search
4. Lead Origin lead add form.
5. What is your current occupation working professional
6. Last Notable Activity unreachable:
   a. modified
   b. email opened
   c. sms sent
7. Last Activity others
8.Total Visits
9. What is your current occupation student
10. What is your current occupation unemployed
11. Last Activity
    a. email opened
    b. sms sent
    c. olark chat conversation

Keeping these in mind the X Education can flourish as they have a very high chance to get almost all the potential buyers to change their mind and buy their courses.

X Education shouldn't focus on the following:

1. Page Views Per Visit
2. Lead Source_reference.