Summary

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

The following steps used:

1. Reading and Understanding the data

Analysis of the data have been done regarding, how much rows and columns are present, how many columns have null values, datatypes of each columns, and basics statiscal knowledge of numerical columns have been checked. There are 9240 rows and 37 columns present, consists of different datatypes.

2. Data Cleaning

This is one of the most crucial part of the assignment in terms of both model buidling and business persceptive and also lots of time consuming.
The data was partially clean except for a few null values and the option select had to be replaced with null values since it did not give us much information. Some of the columns having more than 45% of null values have been removed, while for rest of the null values , some imputation have been done. Here for categorical columns we have used most occuring values for imputation and for numerical columns impute either mean or mode considering the fact that outliers may be present.
Also some columns who are highly skewed or have only one values have been removed while for some columns some categories have been merged in order to give us proper observation for further study.

3. EDA

A quick EDA was done to check the condition of the data. It was found that a lot of elements in the categorical variables were irrelevant.

4. Data Preparation

In this dummy variables have been introduced for the categorical variables, and for numerical variables used MinMaxScaler in order to remove any outliers present in the data. Then we split the data into two parts - Train and Test Data (70% an 30%)
After the following steps , there are 36 columns are present , from which logisitic regression will be build for further prediction.

5. Model Building

First RFE was done to attain 20 relevant variables. Later the rest were removed manually depending on the VIF values and p-values. (The variable with VIF <5 and p-value <0.05 were kept).

6. Model Evaluation

Earlier cutoff value 0.5 was used, from which we get accuracy 86% , and similarly other metrics such as sensitivity - 75.8% and specificity - 92.1% were calculated using confusion metrics.
But are main aim is to increase the leads, which can be achieved by imporving the sensitivity, therefore optimal cutoff point 0.35 is found for building the model, which have given about 84% sensitivity keeping accuracy as 84.5% and specificity 84.8% which is quiet acceptable.
Futher ROC curve also drawn with calculating other metrics such as precision (77.3%) and recall (84.2%) which can be used for future analysis if there is any change in business requirement.

7. Model Prediction

Prediction was done on the test data with an optimal cutoff point 0.3 with sensitivity, accuracy and specificity 84.2%, 82.3% and 84.8% respectively.

It was found that the variables that mattered the most in potential buyers are :

1. TotalVisits
2. Current Occupation (professional)
3. Total time spent on the website
4. Tags - Revert after reading the email
5. Specialization.