Summary

This analysis is done for X Education and to find ways to get more industry 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.

The following are the steps we used:

1. Read and Understand the Data set:

We import all required libraries for this Case study and read the given data and look at the shape of the data set and column details to know how many Numerical and Categorical columns are there.

2. EDA & Cleaning data:
    i. As we can see there are lot of columns which have high number of missing values. Clearly, these columns are not useful. Since, there are 9000+ datapoints in our data set. So, we eliminated the columns having greater than 3000 missing values as they are of no use to us.
    ii. There are a few columns in which there is a level called 'Select' which basically means that the student had not selected the option for that particular column which is why it shows 'Select'. These values are as good as missing values and hence we need to identify the value counts of the level 'Select' in all the columns that it is present.
    iii. Clearly the levels "Lead Profile" and "How did you hear about X Education" have a lot of rows which have the value Select which is of no use to the analysis so we dropped them.
    iv. Also notice that when you got the value counts of all the columns, there were a few columns in which only one value was majorly present for all the data points. These include 'Do Not Call', 'Search', 'Magazine', 'Newspaper Article', 'X Education Forums', 'Newspaper', 'Digital Advertisement', 'Through Recommendations', 'Receive More Updates About Our Courses', 'Update me on Supply Chain Content', 'Get updates on DM Content', 'I agree to pay the amount through cheque'. Since practically all of the values for these variables are No, hence we dropped these columns as they won't help with our analysis.
    v. Since X education is an online education platform considering the variable "City" & "Country" won't be of any use in our analysis. So we dropped it.
    vi. Also, by looking at values in the variables "Prospect ID" and "Lead Number" won't be of any use in the analysis, so we dropped these two variables.
    vii. The variable "What matters most to you in choosing a course" has the level Better Career Prospects 6528 times while the other two levels appear very less. So we dropped this column as well.
    viii. Now, there's the column What is your current occupation which has a lot of null values. Now you can drop the entire row but since we have already lost so many

feature variables, we choose not to drop it as it might turn out to be significant in the analysis.

ix. We removed the rows of the columns where very fewer null values are there.

x. So, after all cleaning operation we got around 69% of the rows which seems good number for our analysis.

3. Dummy Variables:

The dummy variables were created and later on the dummies with 'select' elements were removed. For numeric values we used the MinMaxScaler.

4. Train-Test split:

The split was done at 70% and 30% for train and test data respectively.

5. Model Building:

Firstly, RFE was done to attain the top 15 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).

6. Model Evaluation:

A confusion matrix was made. Later on the optimum cut off value (using ROC curve) was used to find the accuracy, sensitivity and specificity which came to be around 75-80%.

7. Prediction:

Prediction was done on the test data frame and with an optimum cut off as 0.35 with accuracy, sensitivity and specificity of 75-80%.

8. Precision – Recall:

This method was also used to recheck and a cut off of 0.42 was found with Precision around 77% and recall around 78% on the test data frame.

9. Final list of features:

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

   a. TotalVisits
   b. Total Time Spent on Website
   c. Lead Origin_Lead Add Form
   d. Lead Source_Olark Chat
   e. Lead Source_Welingak Website
   f. Do Not Email_Yes
   g. Last Activity_Had a Phone Conversation
   h. Last Activity_SMS Sent
   i. What is your current occupation_Student
   j. What is your current occupation_Unemployed
   k. Last Notable Activity_Unreachable

10. Suggestion:

i. 'Total Visits', 'Total Time Spent on Website', 'Page Views Per Visit' which contribute most towards the probability of a lead getting converted.

ii. The Sales team needs to target the people to below leads.

- The people who all are visiting the site and spending maximum time.
- Connect the people where earlier a phone conversation happened but didn’t enrol for any course yet
- People belongs to category Unemployed/ Student category as they might be looking for any courses to upskill their skill set and looking for Job opportunities out of that.