Summary

This is the summary of analysis done for X Education and to reach out to industry professionals to join their courses. The basic data provided gave us a lot of information about how the potential customers visit per website, how much time they spend there, how they reached the site and the conversion rate.

The following are the steps used:

1. Data Inspection: First start the project with data inspection, where we check the shape of the data, check the null values counts in variables by using info function and describe the data to check the min, max, counts, etc.

2. 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. Dropped few columns who have null values more than 40. Since most values are 'India', we can impute missing values in this column with this value.

3. Exploratory Data Analysis:

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, so dropped those variables. The numeric values seem good.

4. Data Preparation:

Converted some binary variables to 0/1. The dummy variables were created for categorical variables. For numeric values we used the Standard Scaler.

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 20 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

81% each.

7. Prediction:

Prediction was done on the test data frame and with an optimum cut off as 0.34 with accuracy, sensitivity and specificity of around 80%.

8. Precision – Recall:

This method was also used to recheck and a cut off of 0.34 was found with Precision around 79% and recall around 71% on the test data frame.

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

1. Lead Origin_Lead Add Form
2. Lead Source_Welingak Website
3. What is your current occupation_Working Professional
4. Last Activity_SMS Sent
5. Total Time Spent on Website
6. Lead Source_Olark Chat

Recommendation:

- The company should make calls to the leads coming from the lead sources "Welingak Websites" and "Reference" as these are more likely to get converted.
- The company should make calls to the leads who are the "working professionals" as they are more likely to get converted.
- The company should make calls to the leads who spent "more time on the websites" as these are more likely to get converted.
- The company should make calls to the leads coming from the lead sources "Olark Chat" as these are more likely to get converted.
- The company should make calls to the leads whose last activity was SMS Sent as they are more likely to get converted.
- The company should not make calls to the leads whose last activity was "Olark Chat Conversation" as they are not likely to get converted.
- The company should not make calls to the leads whose lead origin is "Landing Page Submission" as they are not likely to get converted.
- The company should not make calls to the leads whose Specialization was "Others" as they are not likely to get converted.
- The company should not make calls to the leads who chose the option of "Do not Email" as "yes" as they are not likely to get converted.