Summary Report

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 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. The columns with more than 40% null values dropped and also some of the columns which are not useful in the analysis dropped.

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 after plotting the count plot. They have been removed from the dataset. Outlier treatment was done for the numerical columns have outliers in it.

3. Data Preparation by Binary Variable Conversion, Dummy Variables Creation and Scaling:

The Before splitting the data into train and test, data preparation step carried out by converting the binary variables into 0/1, dummy variables creation done using the categorical variables. Scaling of numerical variables was done.

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 p-value and VIF (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.

7. Prediction:

Prediction was done on the test data frame and with an optimum cut off as 0.36.

8. Precision – Recall:

This method was also used to recheck and a cut off of 0.41 was found with Precision around 72% and recall around 80% on the test data frame.

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

1. Lead Source_Welingak Website.
2. Lead Source_Reference.
3. What is your current occupation_Working Professional.
4. Last Activity_Other Activity.
5. Last Notable Activity_Unreachable.
6. Last Notable Activity_SMS Sent.
7. Last Activity_Unsubscribed.
8. Lead Source_Olark Chat.
9. Total Time Spent on Website.
10. Lead Origin_Landing Page Submission
11. Specialization_Not Available
12. Last Activity_Olark Chat Conversation
13. Do Not Email

Keeping these in mind the X Education can be very successful as they have a very high chance to get almost all the potential leads to change their mind and join their courses.