Summary

Problem Statement:

An education company named X Education sells online courses to industry professionals. On any given day, many professionals who are interested in the courses land on their website and browse for courses. although X Education gets a lot of leads, its lead conversion rate is very poor. although X Education gets a lot of leads, its lead conversion rate is very poor.

We used following steps to solve the above-mentioned problem:

1. Read and Inspect the data: We read the data and checked the shape, data types, null values and summary

2. Cleaning data: In the given dataset, 17 features had null values and there were many cases where data was provided as “Select”. For cleaning this we replaced all the “Select” will null values. We removed the rows with high null values and rechecked the null value percentage in each column and cleaned it.

3. EDA: We used boxplot, countplot and heatmap for EDA. Using countplot we compared the feature values with respect to converted column. Using boxplot we found that “TotalVisits” and “Page views per visits” have outlier and we cleaned that data and at last we used headmap to check the relation between the features.

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 90% .

7. Prediction: Prediction was done on the test data frame and with an optimum cut off as 0.42 with accuracy, sensitivity and specificity of 90%.

8. Precision – Recall: This method was also used to recheck and a cut off of 0.42 was found with Precision around 89% and recall around 90% on the test data frame. 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. Total number of visits.

3. When the lead source was:

a. Google

b. Direct traffic

c. Organic search

d. Welingak website

4. When the last activity was:

a. SMS

b. Olark chat conversation

5. When the lead origin is Lead add format.

6. When their current occupation is as a working professional. 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.