Summary

Problem Statement:

An education company named X Education sells online courses to industry professionals. On any given day, a large number of professionals who are interested in the courses land on their website and browse for courses. Although X Education receives a lot of leads, its lead conversion rate is extremely poor. We used the following steps to solve the above-mentioned problem:

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

2. Cleaning the data: In the given dataset, 17 features had null values and there were a lot of cases where data was provided as "Select". For the cleaning process we replaced all the "Select" will null values and also removed the rows having high null values and reexamined the null value percentage in each column and cleaned it.

3. EDA: We utilized boxplot, countplot and heatmap for EDA. By using countplot we compared the feature values with respect to the converted column. Using boxplot we found out that the "TotalVisits" and "Page views per visits" have outlier and we cleaned that data and lastly we used heatmap to check the relation between the features.

4. Train-Test split: The split was conducted at 70% and 30% for train and test data respectively.

5. Model Building: Firstly, RFE was done to obtain the top 20 relevant variables. Later the remaining 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 created. Later on the optimum cut off value (using ROC curve) was used to determine the accuracy, sensitivity and specificity which came to be around 81% .

7. Prediction: Prediction was done on the test data frame and with an optimum cut off as 0.41 and also with accuracy, sensitivity and specificity of 80%.

8. Precision – Recall: This method was also used to recheck and a cut off of 0.41 was found with precision around 70% and recall around 80.18% on the test data frame. It was found out that the variables that mattered the most in the potential buyers are (in descending order):

I. The total time spent on the Website.
II. Total number of visits.
III. What the lead source was:
    1) Google
    2) Direct traffic
    3) Organic search

4) Welingak website

4. When the last activity was:

I. SMS
II. Olark chat conversation

5. When the lead origin is Landing Page Submission.

6. When their current occupation is a working professional.

Keeping all this in mind the X Education can flourish as they have a very high probability to get almost all the potential buyers to change their mind and purchase their courses.