Lead Scoring Case Study - Summary

This analysis is done for X Education 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 selected had to be replaced with a null value since it did not give us much information. A few of the null values were changed to 'not given' so as to not lose much data.

We have also dropped some columns -

a. Which have missing values of more than 35%
b. Also a few unwanted columns.

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. The numeric values seem good and A few outliers were found and we treated them accordingly.

3. Dummy Variables:
The dummy variables were created and later on the dummies with 'not given' 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 accuracy, sensitivity, & specificity trade-off)
was used to find the accuracy, sensitivity, and specificity which came to be around 80% each.

7. Prediction:
The prediction was done on the test data frame and with an optimum cut of 0.35 with

accuracy, sensitivity, and specificity of 80%.

The Precision & Recall were around 78% & 75%

8. Precision – Recall:
This method was also used to recheck and a cut-off of 0.40 was found with a Precision around 73% and recall around 75% on the test data frame.

Conclusion:

After the analysis - We can conclude that the model 5 (please refer python file) with 0.35 as a cut off is delivering the Recall value of 78.7 % on training dataset & 80% on the test dataset. This can be considered as a reasonable performance.

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

1)The total time spent on the Website.
2) The total number of visits.

3)When the lead source was:
   a. References
   b. Olark chat conversation
   c. Organic search
   d. Welingak website

4) When the last activity was:
   a. SMS
   b. Had a Phone Conversation

5) When their current occupation is as a working professional.

Keeping these in mind X Education can flourish as they have a very high chance to get almost all the potential buyers to change their minds and buy their courses.