Summary | Lead Scoring Assignment

This analysis is done for X Education company 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 in our analysis:

1. Cleaning data: The data was partially clean except for a few null values and the select values were changed to 'NaN' as they did not provide any useful information. Later they were handled along with the existing NULL values. A few columns were discarded that consisted majorly of NULL values.

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. Hence, they were replaced with "rare" to not loose data and reduce the number of unique categories. The numeric values seemed good and outliers were found and handled in few columns.

3. Dummy Variables: The dummy variables were created for the categorical columns. For numeric values, we used the StandardScaler to scale the data into a more compact version for better utilization.

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

5. Model Building: The heatmap was plotted for all the columns and based on it a few highly correlated and insignificant dummy variables were removed. Then RFE was performed to attain the top 15 relevant variables. Later few more variables were removed one by one manually depending on the VIF values and p-values (The variables with VIF < 5 and p-value < 0.05 were kept).

6. Model Evaluation: A confusion matrix was plotted on the train data. Later on, the optimum cut off value (using ROC curve) was used to find the accuracy, sensitivity and specificity which came to be around 80% each.

7. Prediction: Prediction was done on the test data frame and with an optimum cut off as 0.35 and with accuracy, sensitivity and specificity of around 79%-80%.

8. Precision-Recall: With a cut off of 0.35, the Precision was found to be around 70%. The recall comes to be around 80% on the test data frame. Hence, the model is able to correctly predict 80% of the lead conversions.

It was found that the variables that positively influenced the most in recognizing the potential leads conversions are

1. Lead Origin_lead add form
2. What is your current occupation_working professional
3. What matters most to you in choosing a course_better career prospects
4. Lead Origin_api
5. Total Time Spent on Website

Summary | Lead Scoring Assignment

X Education company should personalize the enquiry experience for working professionals to enhance customer experience and conversions. Leads spending more time on the website can be tracked and targeted with ads based on their search history. Keeping the other major factors in mind and using such strategies can help X Education convert a lot of potential buyers into customers.