SUMMARY

X Education, an online education company, aims to improve its lead conversion rate from 30% to 80% by identifying 'Hot Leads' using a logistic regression model. They have a dataset with 9,240 entries and 37 attributes, including Lead Source, Total Time Spent on Website, and Last Activity, with target variable 'Converted' indicating whether a lead converted (1) or not (0). The dataset also contains categorical variables with placeholders like 'Select' for missing values.

The objective is to build a logistic regression model to assign lead scores from 0 to 100, prioritizing leads with a higher likelihood of conversion and to adapt the model to future changes in company requirements and adjust solutions based on evolving needs. The final goal is to use the model to enhance lead scoring, thereby improving conversion rates and overall efficiency in turning leads into customers.

From the problem statement, many of the categorical variables have a level called 'Select' which needs to be handled because it seems as Null value. Hence convert 'Select' to NaN. Calculate the proportion of null values for each column and remove columns with over 40% missing values.

The city column showed NaN as most occurring city, with null values replaced with mode i.e. Mumbai since it is a categorical column. Specializations that were not listed were marked as "Not Specified" to maintain data integrity, and Not Specified has the highest number of leads and second place Business_and_Management. What matters most to you in choosing a course? Showed results Better Career Prospects with 6528, NaN with 2709, Flexibility & Convenience with 2 and Others with 1 value counts. Replaced the NaN with mode i.e. 'Better Career Prospects' since it is categorical column. What is your current occupation? Showed results Unemployed 5600, NaN with 2690, Working Professional with 706, Student with 210, Other with 16, Housewife with 10, Businessmen with 8. Replaced the NaN with mode i.e. 'Unemployed' since it is a categorical column. However, to prevent bias in analysis, the country column was removed as India's prevalence could skew results.

Total Visits shown 9103 with mean 3.44. Page visits per visit with mean 2.0. Replaced lower frequency values with new category "Others" for "Last Activity". Lead sources like Google and Direct traffic yielded the most leads with 2836 and 2499 respectively. Clubbed similar spelling and common attributes Google and google and Facebook and Social Media are similar terms into single. Replacing lower frequency values with new category "Others" for "Lead Source".

With threshold value of 0.5, Model has accuracy with 92.819%, Sensitivity with 88.53%, Specificity with 95.508%. The area under ROC curve is 0.97 which is good. With the optimal threshold value of 0.29, Test Model has accuracy with 92.367%, sensitivity with 92.959%, specificity with 92.028%. Customers with a Lead score of 80% or higher can be treated as “Good Leads” and can be approached for positive conversion. Top 3 features contributing most towards lead conversion: Tags_Closed by Horizzon, Tags_Lost to EINS, Tags_Will revert after reading the email.