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 websites and browse for courses.

The company markets its courses on several websites and search engines like Google. Once these people land on the website, they might browse the courses, fill up a course form, or watch some videos. When these people fill out a form providing their email address or phone number, they are classified as a lead. Moreover, the company also gets leads through past referrals. Once these leads are acquired, employees from the sales team start making calls, writing emails, etc. Through this process, some of the leads get converted while most do not.

Overall Approach to Solution:

Reading and Understanding data

Imported the dataset and check the basics of the dataset to get an understanding of the data initially. [shape, describe, info]

Data Cleaning and Preparation

- Checked for NULL values
- Converted the 'Select' values in certain variables to NaN
- Treated all the NULL values o If the NULL values in the column exceeded 45%, dropped the column o imputing – Mean and Median [Numerical data], Mode [Categorical data]
- Dropped the rows if they were insignificant
- Combined some categories in certain variables to a single entity as they did not show much difference alone.
- Removed variables where only 1 value dominated
- Visualized the Numerical and Categorical data and made observations.

Preparing the data for Modelling

Created dummy variables for the categorical data.

Train-Test Split

The Split the data into training (70%) and testing (30%) sets to evaluate the model's performance.

Scaled the features using StandardScaler().

Created dummy variables for Categorical variables.

Data Visualization:

Created visualizations such as count plots and box plots to explore the distribution of data and relationships between variables.

Explored the relationships between variables using pair plots and heatmaps.

Model Evaluation

Now, prediction on the **train set.**

Accuracy –84.29 %, Sensitivity – 83.69 %, Specificity – 84.66 %

- Checked for the area under the ROC curve as a metric to evaluate the model – 0.92
- Found the optimal threshold - to be 0.327

Prediction on the Test set

- Scaled the test dataset using the StandardScaler () as in the train set [expect this time only transform was done and not fit]

Accuracy – 82.51 %,
Sensitivity – 85.44%,
Specificity – 80.85%

Insights

The key determinants that impact the likelihood of a lead converting into a customer include:

Lead Origin: Specifically, leads generated through 'Lead Add Form.'
Current Occupation: Particularly, leads categorized as 'Working Professionals.'
Lead Source: Notably, leads originating from 'Welingak Website.'
These attributes serve as valuable insights for optimizing targeted marketing campaigns. For instance, when a company seeks to market a product or service to individuals in the working professional segment, they should concentrate their marketing endeavors on leads that possess these specific attributes.

Here are the strategies to consider:

The sales team should give priority to leads that have originated from 'Lead Add Forms.'
Special attention should be directed toward leads sourced from the 'Welingak Website.'
A primary focus should be on engaging with individuals identified as 'Working Professionals.'

Although students can be approached, it's important to acknowledge that their probability of conversion might be lower, as they are currently engaged in their studies and may be less inclined to enroll in a course tailored for working professionals.