SUMMARY

This assessment is conducted for X Education to discover methods for attracting more industry professionals to enroll in their courses. The initial data provided offered significant insights into how potential customers navigate the site, the duration of their visits, the channels through which they accessed the site, and the conversion rate.

Below are the steps implemented:

1. Data Cleaning:
   The dataset was mostly clean, but a few null values were present, and the option select was replaced with a null value as it provided little information. Some null values were transformed to ‘not provided’ to retain more data, although they were eventually dropped when creating dummies. Given the abundance of entries from India compared to those from abroad, the categories were updated to ‘India,’ ‘Outside India,’ and ‘not provided.’
2. Exploratory Data Analysis (EDA):
   A brief EDA was conducted to assess the state of the data. It was discovered that several elements in the categorical variables were not relevant. The numeric values appeared satisfactory, and no outliers were detected.
3. Dummy Variable Creation:
   Dummy variables were created, and subsequently, those associated with ‘not provided’ were eliminated. For the numeric values, the MinMaxScaler was applied.
4. Splitting the Data:
   The dataset was split into 70% for training and 30% for testing.
5. Model Development:
   Initially, Recursive Feature Elimination (RFE) was performed to identify the top 15 significant variables. The remaining variables were then manually excluded based on their Variance Inflation Factor (VIF) and p-values (variables with VIF < 5 and p-value < 0.05 were retained).
6. Model Assessment:
   A confusion matrix was created. The optimal cut-off value, determined using the ROC curve, allowed for calculating accuracy, sensitivity, and specificity, all approximately 80%.
7. Prediction:
   Predictions were made on the test dataset, employing an optimal cut-off of 0.35, leading to accuracy, sensitivity, and specificity values of 80%.
8. Precision-Recall Analysis:
   This technique was also utilized to verify results, revealing a cut-off of 0.41, with precision around 73% and recall approximately 75% on the test dataset.

Study indicated that the most significant factors influencing potential buyers are (in order of importance):

1. The total duration spent on the website.

2. The overall number of visits.

3. The lead source being:

    a. Google

    b. Direct traffic

    c. Organic search

    d. Welingak website

4. The timing of the last activity being: a. SMS b. Olark chat interaction

5. The lead origin being in Lead add format.

6. The current occupation of the individual as a working professional.

With these insights, X Education can thrive, as they have a strong likelihood of persuading nearly all potential buyers to reconsider and enroll in their courses.