SUMMARY

The analysis aims to assist X Education in increasing enrolment of industry professionals in their courses. The initial dataset offers valuable insights into visitor behaviour, including their website interactions, session durations, referral sources, and conversion rates.

The approach involved the following stages:

1. Data Cleaning:

The dataset was moderately clean, except for a handful of missing values. The "select" option was substituted with null values due to its limited informational value. A small portion of the missing records are deleted as it would not make any significant difference. In light of the distribution of respondents—comprising a significant number from India and a minority from other regions, we dropped the country column as it shows data imbalance. Also, the "City" column was dropped as most of the people chose select option which doesn't signify anything.

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 no outliers were found.

3. Dummy Variables:

The dummy variables were created for categorical variables. 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 20 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. Assessing the Model:

We generated a confusion matrix to evaluate the model's performance. Subsequently, leveraging the ROC curve, we determined the optimal threshold value. This helped us compute accuracy, sensitivity, and specificity, which was approximately 80%.

7. Making Predictions:

Predictions were applied to the test dataset using an optimal threshold of 0.43. This yielded an accuracy, sensitivity, and specificity of almost 80%.

8. Precision and Recall Analysis:

Additionally, we employed the precision-recall approach for further verification. By selecting a threshold of 0.43, we achieved a precision of around 78.8% and a recall of approximately 77.6% on the test dataset.

To enhance conversion rates, implement the following strategies:

- Concentrate efforts on leads from "Welingak Website" and "Reference" sources, known for higher conversion potential.
- Prioritize working professional leads, as they exhibit a stronger inclination to convert.
- Engage leads who spent more time on the website, as their active involvement suggests increased conversion likelihood.
- Pay special attention to "Olark Chat" leads, which have a higher chance of conversion.
- Utilize "SMS Sent" activity by reaching out to leads with this status for better conversion prospects.
- Consider avoiding leads with "Olark Chat Conversation" activity, which may have lower conversion rates.
- Exercise caution with "Landing Page Submission" leads due to their historically lower conversions.
- Selectively approach leads with "Others" specialization, given their lower conversion probability.
- Minimize contact with "Do Not Email" leads, as their conversion likelihood.