Summary

This analysis was conducted for X Education to find ways to attract more industry professionals to their courses. The provided dataset offered insights into how potential customers visit the site, the time they spend there, how they arrived, and the conversion rate.

Technical Steps Used:

1. Data Cleaning:

- Removed redundant variables/features from the dataset.
- Replaced the option 'Select' with null values as it provided little information.
- Dropped columns with more than 40% null values.
- Checked the number of unique categories for all categorical columns.
- Identified and dropped highly skewed columns.
- Treated missing values by imputing with mean, median, or mode as appropriate.
- Detected and addressed outliers.

2. Exploratory Data Analysis:

- Conducted a quick EDA to assess data condition, revealing irrelevant elements in categorical variables and outliers in numeric values.
- Performed univariate analysis for both continuous and categorical variables.
- Conducted bivariate analysis with respect to the target variable.

3. Dummy Variables:

- Created dummy variables for all categorical columns.

4. Scaling:

- Scaled the data for continuous variables using a standard scaler.

5. Train-Test Split:

- Split the data into 70% for training and 30% for testing.

6. Model Building:

- Used Recursive Feature Elimination (RFE) to select the top 20 relevant variables.

- Removed irrelevant features manually based on Variance Inflation Factor (VIF) values and p-values, retaining variables with VIF < 5 and p-value < 0.05.

7. Model Evaluation:

- Created a confusion matrix and used the ROC curve to determine the optimal cut-off value, achieving an accuracy, sensitivity, and specificity of around 80%.

8. Prediction:

- Made predictions on the test data with an optimal cut-off of 0.37, achieving an accuracy, sensitivity, and specificity of 80%.

9. Precision-Recall:

- Rechecked using precision-recall analysis with a cut-off of 0.41.

Conclusion:

The most important variables for identifying potential buyers are:

- Total time spent on the website.

- Total number of visits.

- Lead source, particularly Olark Chat.

- Last activity, especially SMS and Olark Chat conversation.