SUMMARY:

PROBLEM STATEMENT:

X Education, an online course provider, seeks assistance in identifying the most promising leads with the highest likelihood of converting into paying customers. The objective is to assign lead scores to prioritize leads, ensuring those with higher scores are more likely to convert, aligning with the CEO's target conversion rate of around 80%.

SOLUTION SUMMARY:

Step 1: Reading and Understanding Data:

The dataset was reviewed to gain insights.

Step 2: Data Cleaning:

a. Variables with unique values were dropped.
b. Entries marked as 'Select' were replaced with null values.
c. Columns with null values exceeding 35% were removed.
d. Imbalanced and redundant variables were addressed, including imputation of missing values and fixing label inconsistencies.
e. Sales team-generated variables were excluded.

Step 3: Data Transformation:

Binary variables were standardized to '0' and '1'.

Step 4: Dummy Variables Creation:

Dummy variables were created for categorical variables, eliminating duplicates.

Step 5: Test Train Split:

The dataset was divided into training and testing sets with a 70-30 split.

Step 6: Feature Rescaling:

a. Min Max Scaling was applied to numerical variables.
b. Correlation analysis was conducted via a heatmap, and highly correlated dummy variables were removed.

Step 7: Model Building:

a. Recursive Feature Elimination was used to select the top 15 important features.
b. Significance of features was determined based on P-values, with insignificant variables dropped.
c. The final model comprised 11 significant variables with satisfactory VIF scores.
d. Optimal probability cutoff was determined through ROC curve analysis, yielding an area under the curve of 86%.
e. Model accuracy, sensitivity, and specificity were assessed, achieving satisfactory results.
f. Precision-recall trade-off led to a cutoff value of approximately 0.3.
g. Learnings from the train set were applied to the test set, resulting in a conversion probability accuracy of 77.52%, sensitivity of 83.01%, and specificity of 74.13%.

Step 8: Conclusion:

- The final predicted model achieved a conversion rate of 83% on the test set, meeting the CEO's target.
- The model's good sensitivity ensures the selection of promising leads.
- Key features contributing to lead conversion probability include –
  Lead Origin_Lead Add Form.
  Working Professional occupation.
  Total Time Spent on Website.