Lead Scoring Case Study Summary

1. Problem Statement:

X Education, an online course provider for industry professionals, seeks assistance in identifying the most promising leads likely to convert into paying customers. The objective is to develop a lead scoring model that assigns lead scores, ensuring higher conversion chances for leads with higher scores and lower chances for those with lower scores. The CEO has set a target lead conversion rate of approximately 80%.

2. Solution Summary:

Step1: Reading and Understanding Data:

The dataset was reviewed and inspected to understand its structure and content.

Step2: Data Cleaning:

a. Identified and dropped variables with unique values.

b. Replaced 'Select' values with null values in columns where leads did not choose any option.

c. Removed columns with more than 70% NA values and addressed imbalanced and redundant variables.

d. Imputed missing values, handled outliers, and standardized labels for consistency. For instance, standardized city names and imputed missing values based on the predominant location (Mumbai in this case).

Step3: Data Transformation:

Converted binary variables to '0' and '1' and transformed the target variable to indicate lead conversion status (1 for converted, 0 for not converted).

Step4: Dummy Variables Creation:

Created dummy variables for categorical variables and eliminated repeated or redundant variables.

Step5: Test Train Split:

Partitioned the dataset into training and testing sets with a 70-30 split.

Step6: Feature Rescaling:

a. Applied Min-Max Scaling to normalize numerical variables.

b. Examined variable correlations via heatmap and removed highly correlated dummy variables.

Step7: Model Building:

a. Employed Recursive Feature Elimination to select the 15 most important features.

b. Utilized statistical analysis, including P-values, to identify significant variables and exclude insignificant ones.

c. Arrived at 13 significant variables with acceptable VIFs.

d. Determined optimal probability cut-off for the final model by assessing accuracy, sensitivity, and specificity.

e. Evaluated model performance using ROC curve analysis, achieving an area under the curve of 89%.

f. Ensured 80% prediction accuracy based on the converted column.

g. Assessed precision, recall, accuracy, sensitivity, and specificity on the train set.

h. Determined a cut-off value of approximately 0.3 based on Precision and Recall trade-off.

i. Implemented learnings on the test model and calculated conversion probability with Sensitivity and Specificity metrics, achieving 89.98% accuracy, 81.79% sensitivity, and 94.95% specificity.

Step 8: Conclusion:

- The lead scoring model predicted a conversion rate of 89.37% on the test set, aligning with the CEO's target rate of around 80%.
- The model's high sensitivity facilitates the identification of promising leads.
- Key features contributing to conversion probability include Lead Origin (Lead Add Form), Current Occupation (Working Professional), and Total Time Spent on Website.