Lead Scoring Case Study Summary

Problem Statement:

X Education sells online courses to industry professionals and needs assistance in identifying the most promising leads, i.e., those most likely to convert into paying customers. The company requires a model that assigns a lead score to each lead, with higher scores indicating a higher chance of conversion and lower scores indicating a lower chance. The CEO has set a target lead conversion rate of around 80%.

Solution Summary:

Step 1: Reading and Understanding Data

- Read and analyze the data.

Step 2: Data Cleaning

- Dropped variables with a high percentage of NULL values.
- Imputed missing values with median values for numerical variables and created new classification variables for categorical variables.
- Identified and removed outliers.

Step 3: Data Analysis

- Conducted Exploratory Data Analysis (EDA) to understand the data orientation.
- Identified and dropped variables with only one value across all rows.

Step 4: Creating Dummy Variables

- Created dummy data for categorical variables.

Step 5: Test-Train Split

- Divided the dataset into training and testing sections with a 70-30% split.

Step 6: Feature Rescaling

- Applied Min-Max Scaling to the original numerical variables.
- Created an initial model using stats models to get a statistical view of all parameters.

Step 7: Feature Selection using RFE

- Used Recursive Feature Elimination (RFE) to select the top 20 important features.
- Evaluated P-values to select the most significant features, resulting in the identification of 15 significant variables with acceptable VIF values.

- Created a data frame with converted probability values, assuming a probability value greater than 0.5 indicates conversion (1) and less than 0.5 indicates no conversion (0).
- Derived Confusion Metrics and calculated the model's overall accuracy.
- Calculated 'Sensitivity' and 'Specificity' to assess model reliability.

Step 8: Plotting the ROC Curve

- Plotted the ROC curve, which showed an area under the curve (AUC) of 89%, indicating a strong model.

Step 9: Finding the Optimal Cutoff Point

- Plotted the probability graph for 'Accuracy,' 'Sensitivity,' and 'Specificity' for different probability values.
- Determined the optimal probability cutoff point at 0.37.
- Found that the model correctly predicted approximately 80% of the values, with accuracy at 81%, sensitivity at 79.8%, and specificity at 81.9%.
- Calculated the lead score, achieving a target lead prediction of around 80%.

Step 10: Computing Precision and Recall Metrics

- Calculated Precision and Recall metrics, resulting in values of 79% and 70.5%, respectively, for the training dataset.
- Based on the Precision-Recall tradeoff, identified a cutoff value of approximately 0.42.

Step 11: Making Predictions on Test Set

- Applied the model to the test dataset and calculated conversion probability using Sensitivity and Specificity metrics.
- Found the accuracy to be 80.8%, sensitivity at 78.5%, and specificity at 82.2%.