Lead Scoring Case Study Summary

Problem Statement:

X Education aims to boost lead conversion rates by identifying the most promising leads. The goal is to create a model that assigns lead scores, indicating higher conversion chances for leads with higher scores. The CEO has set a target lead conversion rate of approximately 80%.

Solution Summary:

Step 1: Reading and Understanding Data

1. Analyzed the dataset to understand its structure and content.

Step 2: Data Cleaning

1. Dropped variables with high NULL percentages.
2. Imputed missing values and handled outliers.

Step 3: Data Analysis

1. Conducted Exploratory Data Analysis (EDA) and dropped variables with a single value across all rows.

Step 4: Creating Dummy Variables

1. Converted categorical variables into dummy variables.

Step 5: Test Train Split

1. Divided the dataset into training and testing sets in a 70-30 proportion.

Step 6: Feature Rescaling

1. Applied Standard Scaling to numerical variables.
2. Created the initial statistical model using the stats Logistic model.

Step 7: Feature Selection using RFE

1. Utilized Recursive Feature Elimination (RFE) to select the top 15 important features.
2. Recursively analyzed P-values for feature significance and retained the most significant ones.

Step 8: Confusion Metrics and Model Accuracy

1. Derived the Confusion Matrix using an assumed probability cutoff of 0.5.
2. Calculated overall model accuracy, sensitivity, and specificity matrices.

Step 9: Plotting the ROC Curve

1. Plotted the Receiver Operating Characteristic (ROC) curve, achieving an area under the curve of 90%.

Step 10: Finding the Optimal Cutoff Point

1. Plotted probability graphs for accuracy, sensitivity, and specificity at different probability values.
2. Determined the optimal probability cutoff point at 0.34, yielding an 81.72% prediction accuracy, Sensitivity= 79.9% and Specificity= 82.15%.

Step 11: Computing Precision and Recall Metrics

1. Calculated precision and recall metrics, resulting in values of 79.6% and 68.9% on the train dataset.
2. Derived a cutoff value of approximately 0.4 based on the Precision and Recall tradeoff.

Step 12: Making Predictions on Test Set

1. Applied learned insights to the test model.
2. Calculated conversion probability using Sensitivity and Specificity metrics.
3. Achieved an accuracy value of 79.83%, with Sensitivity at 75.99% and Specificity at 82.14%.

Conclusion:

The comprehensive approach, encompassing data cleaning, feature selection, and model evaluation, resulted in a robust logistic regression model for lead scoring. The model's performance metrics align with the CEO's target, providing actionable insights for optimizing lead conversion strategies.