Lead Scoring Analysis for X Education

Objective

X Education, a provider of online courses for industry professionals, aimed to enhance their lead conversion strategy by identifying the most promising leads likely to convert into paying customers. The company required a model to assign a lead score, ensuring an 80% lead conversion rate.

Methodology

1. **Data Preparation**
   - **Data Cleaning**:
     - Addressed missing values by replacing them with median values for numerical data and creating new classifications for categorical variables.
     - High-NULL-value variables and irrelevant features (e.g., single-value variables) were dropped.
   - **Handling Outliers**:
     - Identified and removed outliers to ensure data integrity.
2. **Exploratory Data Analysis (EDA)**
   - Analyzed categorical and numerical variables to assess their relevance and quality.
   - Checked for outliers and distribution patterns, confirming the dataset's readiness for modeling.
3. **Feature Engineering**
   - Created dummy variables for categorical data.
   - Applied Min-Max Scaling to normalize numerical values.
4. **Train-Test Split**
   - Divided the dataset into training (70%) and testing (30%) sets to evaluate the model's generalization.
5. **Feature Selection**
   - **Recursive Feature Elimination (RFE)**: Selected the top 15 variables with high predictive power.
   - Retained features based on p-values (<0.05) and Variance Inflation Factor (VIF < 5).

Model Development

1. **Initial Model**
   - Built the initial logistic regression model to assess statistical parameters.
2. **Performance Metrics**
   - Generated a confusion matrix to evaluate accuracy, sensitivity, and specificity.
   - Achieved approximately 80.35% for all three metrics using an optimal cutoff value of 0.5.
3. **ROC Curve Analysis**
   - The ROC curve exhibited an 88% area under the curve (AUC), indicating strong model performance.
4. **Optimal Cutoff Selection**
   - Determined the ideal probability threshold (0.39) by analyzing intersections of accuracy, sensitivity, and specificity curves.

Evaluation and Insights

1. **Performance Metrics**
    - Achieved final metrics:
        - **Accuracy**: 80.73%
        - **Sensitivity**: 80%
        - **Specificity**: 81%
    - Precision and recall values of 79% and 67%, respectively, with an adjusted cutoff of 0.4 for optimal trade-off.
