Lead Scoring Case Study Summary

Problem Statement

X Education, an organization specializing in online courses for industry professionals, aims to improve its lead conversion process. The company needs a data-driven approach to identify high-potential leads—those most likely to convert into paying customers.

To achieve this, a **lead scoring model** is required, assigning a numerical score to each lead. Higher scores indicate a greater likelihood of conversion, while lower scores suggest a reduced probability. The CEO has set a **target lead conversion rate of approximately 80%**, making it crucial to optimize the selection of potential customers.

Solution Approach

Step 1: Data Collection and Understanding

The initial phase involved gathering, exploring, and analyzing the dataset to understand its structure, distribution, and key attributes affecting lead conversion.

Step 2: Data Cleaning and Preprocessing

- **Handling Missing Values**: Variables with a high percentage of missing values were dropped.
- **Imputation**: Missing numerical values were replaced with **median values**, while categorical variables were adjusted by creating appropriate classifications.
- **Outlier Treatment**: Outliers were identified and removed to improve model robustness.

Step 3: Exploratory Data Analysis (EDA)

- Conducted **EDA** to assess the distribution and relationships among variables.
- Three variables with **constant values across all rows** were identified and dropped due to their lack of contribution to the model.

Step 4: Encoding Categorical Variables

- Created **dummy variables** for categorical features to ensure compatibility with the model.

Step 5: Splitting the Data into Training and Testing Sets

- The dataset was divided into **70% training data** and **30% testing data** to build and validate the model.

Step 6: Feature Scaling

- **Min-Max Scaling** was applied to standardize numerical features.
- A **statistical model** was created to analyze the impact of each feature on lead conversion.

Step 7: Feature Selection Using Recursive Feature Elimination (RFE)

- **Recursive Feature Elimination (RFE)** was employed to identify the most significant predictors.
- A combination of **P-values and Variance Inflation Factor (VIF)** was used to refine feature selection.
- The final model retained **15 key features** that had the highest influence on conversion probability.
- The probability of conversion was initially assumed to be **1 if greater than 0.5, otherwise 0**.
- **Confusion Metrics** were calculated to assess overall model performance, including accuracy, sensitivity, and specificity.

Step 8: Plotting the ROC Curve

- The **ROC (Receiver Operating Characteristic) curve** was generated to evaluate the model's predictive power.
- The **AUC (Area Under the Curve) score was 95%**, indicating a strong classification ability.

Step 9: Determining the Optimal Cutoff Point

- Plotted probability curves for **accuracy, sensitivity, and specificity** to determine the best cutoff value.
- The **optimal cutoff probability** was found to be **0.3**, improving prediction accuracy.
- Final performance metrics at this cutoff:
    - **Accuracy**: 87.8%
    - **Sensitivity (Recall)**: 88.1%
    - **Specificity**: 87.6%
- The lead scoring system successfully predicted conversions at a rate **close to the 80% target** set by the CEO.

Step 10: Evaluating Precision and Recall Tradeoff

- **Precision** and **Recall** metrics were calculated to assess the tradeoff between false positives and false negatives.
- The values obtained:
    - **Precision**: 91.6%
    - **Recall**: 81.1%
- The optimal cutoff value based on **Precision-Recall balance** was determined to be **0.35**.

Step 11: Model Validation on the Test Set

- The trained model was applied to the test dataset.
- Conversion probability was evaluated using **Sensitivity and Specificity** metrics.
- Final test set performance:
    - **Accuracy**: 87%
    - **Sensitivity**: 84.3%
    - **Specificity**: 82.2%

Conclusion

The developed lead scoring model provides **a highly accurate approach** to identifying promising leads, enabling X Education to optimize its sales process. By leveraging a **data-driven strategy**, the company can efficiently prioritize high-potential leads, ultimately improving conversion rates and aligning with the CEO's 80% target.

✓ Effective feature selection improved model accuracy.
✓ Optimized probability cutoffs ensured a balance between precision and recall.
✓ The model provides actionable insights to enhance lead targeting and conversion strategies.