Lead Score Case Study Summary

Problem Statement: X Education sells online courses to industry professionals. X Education needs help in selecting the most promising leads, i.e. the leads that are most likely to convert into paying customers. The company needs a model wherein you a lead score is assigned to each of the leads such that the customers with higher lead score have a higher conversion chance and the customers with lower lead score have a lower conversion chance. The CEO, in particular, has given a ballpark of the target lead conversion rate to be around 80%

Solution Summary:

Solution summary outlines a structured approach to solving X Education's problem of lead scoring. It provides a clear overview of the steps involved in data preprocessing, feature selection, model building, and evaluation. Here's a breakdown of the key steps:

Data Exploration and Preprocessing:

- Our initial steps involved loading and understanding the dataset.
- To clean the data, we removed variables with a high percentage of NULL values and addressed missing data by imputing numerical variables with median values. This process led to the creation of new categorical variables.
- Outliers were identified and subsequently eliminated to ensure data quality.

Data Analysis:

- An exploratory data analysis was performed to gain insights into the dataset's characteristics.
- Variables that held constant values across all rows were dropped, as they did not provide meaningful information.

Feature Engineering:

- Categorical variables were encoded as dummy variables, making them more suitable for modeling.

Data Splitting:

- We divided the dataset into training and testing subsets, allocating 70% to training and 30% to testing.

Feature Scaling:

- We applied Min-Max Scaling to normalize numerical variables, ensuring consistent scaling across the dataset.
- An initial model was built using statistical techniques to analyze model parameters.

Feature Selection with RFE:

- Recursive Feature Elimination (RFE) was utilized to identify the top 20 most significant features.

- These features were selected based on their statistical significance (P-values), and less influential features were removed. Subsequently, a VIF (Variance Inflation Factor) analysis confirmed their validity.

Probability Calculation and Model Evaluation:

- We calculated binary conversion probabilities, considering a threshold of 0.5.
- The model's performance was assessed using key metrics, including accuracy, Sensitivity, and Specificity.

ROC Curve Analysis:

- To visually assess model performance, we plotted the Receiver Operating Characteristic (ROC) curve.
- The curve demonstrated strong performance, with an impressive area coverage of 97%, affirming the model's effectiveness.

Optimal Cutoff Determination:

- Probability graphs were created for 'Accuracy,' 'Sensitivity,' and 'Specificity' by varying the probability threshold.
- The optimal probability cutoff point was identified at 0.3, leading to enhanced accuracy (92.7%), sensitivity (91.6%), and specificity (93.4%).

Testing on the Holdout Set:

- The insights and optimal cutoff were applied to the test dataset, allowing us to calculate conversion probability.
- The model displayed remarkable performance on the test set, achieving an accuracy of 91.5%, sensitivity of 89.6%, and specificity of 92.6%.

This detailed summary outlines the systematic steps taken to develop a lead scoring model. It showcases robust data preprocessing, effective feature selection, and a model with high predictive accuracy, sensitivity, and specificity, all aligned with the initial problem statement.