Summary of Lead Scoring Case Study:

Problem Statement:

X Education, an online course provider, seeks assistance in identifying the most promising leads that are likely to convert into paying customers. The objective is to assign a lead score to each lead, with higher scores indicating a higher chance of conversion and lower scores indicating a lower chance of conversion. The target lead conversion rate set by the CEO is around 80%.

Solution Summary:

Step 1: Data Analysis and Understanding:

The data was analyzed to gain insights into the variables.

Step 2: Data Cleaning:

Variables with a high percentage of NULL values were dropped. Missing values were imputed using median values for numerical variables and new classification variables were created for categorical variables. Outliers were identified and removed.

Step 3: Exploratory Data Analysis (EDA):

EDA was performed to understand the data distribution. Variables with only one value in all rows were identified and dropped.

Step 4: Creating Dummy Variables:

Dummy variables were created for categorical variables.

Step 5: Test-Train Split:

The dataset was divided into training and testing sets with a 70-30% proportion.

Step 6: Feature Rescaling:

Min-Max scaling was applied to rescale the numerical variables. An initial model was created using stats model to obtain a statistical view of all model parameters.

Step 7: Feature Selection using RFE:

By utilizing Recursive Feature Elimination, we proceeded to identify the top 20 important features. Through the analysis of generated statistics, we recursively examined the p-values to determine the presence of the most significant values and discarded the insignificant ones. As a result, we narrowed down our selection to the 15 most significant variables, which exhibited good Variance Inflation Factor (VIF) values.

Subsequently, we constructed a data frame containing the converted probability values. Initially, we assumed that a probability value exceeding 0.5 corresponds to a classification of 1, while values below 0.5 indicate a classification of 0. Based on this assumption, we derived the Confusion Metrics and calculated the overall Accuracy of the model. Additionally, we computed the Sensitivity and Specificity matrices to assess the reliability of the model.

Step 8: ROC Curve:

The ROC (Receiver Operating Characteristic) curve was plotted for the selected features, indicating a decent area coverage of 89% and further confirming the model's performance.

Step 9: Optimal Cutoff Point:

The probability graph for accuracy, sensitivity, and specificity was plotted for different probability values. The intersecting point of these graphs was determined as the optimal cutoff point, which was found to be 0.37. Around 80% of the values were correctly predicted based on this cutoff point. The accuracy, sensitivity, and specificity were calculated as follows: accuracy=81%, sensitivity=79.8%, specificity=81.9%.

Step 10: Precision and Recall Metrics:

Precision and Recall metrics were computed on the train dataset, yielding values of 79% and 70.5% respectively. The tradeoff between Precision and Recall led to a cutoff value of approximately 0.42.

Step 11: Predictions on Test Set:

The model learnings were applied to the test set, and conversion probabilities were calculated based on Sensitivity and Specificity metrics. The accuracy value was determined as 80.8%, with Sensitivity of 78.5% and Specificity of 82.2%.