SUMMARY REPORT

To address the problem presented by X Education Company, we executed a comprehensive Logistic Regression analysis on the provided dataset. Our approach was structured to generate accurate scores for each lead. We began by loading the data into a Jupyter notebook. The first task was to perform missing value treatment for all columns, addressing any outliers in numeric columns. We excluded columns that had more than 70% missing data to maintain the integrity of our analysis. Next, we conducted Exploratory Data Analysis (EDA) on each categorical column to identify and visualize data imbalances, subsequently dropping columns with high data imbalance.

After this initial data preparation, we split the dataset into training and testing sets using the sklearn library. The training data was then scaled using the Standard Scaler to ensure uniformity. We employed Recursive Feature Elimination (RFE) on the scaled data for coarse tuning, which helped us select 15 key features. Logistic Regression was performed using the statsmodels GLM method. This was followed by an assessment of the coefficients, p-values, and Variance Inflation Factor (VIF) for the selected features. High p-value features were iteratively dropped through two rounds to refine the model further.

In our final model, predictions were made based on the selected features. We assumed a probability cutoff of 0.5 for lead conversion, classifying a lead as converted if the predicted probability exceeded this threshold. This approach enabled us to construct a confusion matrix, achieving an accuracy of 92%, a specificity of 96%, and a sensitivity of 86%.

To evaluate the model's robustness, we plotted the Receiver Operating Characteristic (ROC) curve, which allowed us to analyze the balance between the True Positive Rate (TPR) and the False Positive Rate (FPR). The ROC curve, generated by varying the probability cutoff from 0.0 to 1.0, demonstrated a 96% area under the curve, indicating a strong model. To further balance sensitivity and specificity, we created confusion matrices for probability thresholds ranging from 0.0 to 0.9, at intervals of 0.1. By plotting accuracy, sensitivity, and specificity for each threshold, we identified an optimal cutoff at 0.2. Reevaluating the model with this new cutoff, we observed an accuracy of 92%, a specificity of 94%, and a sensitivity of 87%.

We then scaled the test data and performed lead conversion predictions on this set, maintaining consistent performance metrics: an accuracy of 92%, a specificity of 94%, and a sensitivity of 87%. This affirmed the model's robustness on unseen data. Finally, we created a lead conversion score, calculated as (conversion probability * 100), to rank leads on a scale from 0 to 100, where higher values indicated a higher likelihood of conversion.

This assignment provided valuable learning opportunities, such as handling missing values and outliers, creating dummy variables for categorical data, utilizing Python libraries for logistic regression, selecting the best model based on balanced sensitivity and specificity, and fostering effective teamwork to solve complex problems.