Summary Report
Lead Scoring Case Study

Problem: X Education wants to build a model where they assign a lead score to each lead such that the

customers with a higher lead score have a higher conversion probability. The business

requirement is to increase the lead conversion rate to around 80%.

Solution Approach:

1. **Data cleaning** involved handling various scenarios, such as dropping columns with high null values (>50%), replacing null values with 'Not Provided' or 'Others' for significant columns, and dropping columns with data imbalances like Country.
2. **EDA** was conducted on the cleaned data, including univariate analysis of categorical and numerical variables, bivariate analysis with the 'Converted' target variable, combining less significant categories based on graphical analysis, and treating outliers using the 1.5 IQR method.
3. **Data pre-processing** steps included converting binary variables (Yes/No) to 1/0, creating N-1 dummy columns for N categories in each categorical column, splitting data into training and test datasets (70:30 ratio), and performing feature scaling on continuous variables.
4. **Logistic Regression model building** involved using RFE to select the top 15 relevant variables, iteratively eliminating variables with VIF > 5 or p-value > 0.05, and rebuilding the model at each stage.
5. **Model evaluation** included obtaining predicted values on the training dataset using a cutoff of 0.5, creating a confusion matrix to calculate accuracy (92%), sensitivity (86%), and specificity (95%). ROC curve was plotted and optimal cut off was calculated to be around 0.2. Accuracy (92%), sensitivity (88%), and specificity (94%) were re-evaluated and

Precision-Recall trade-off observed.

6. **Predictions** were made on the test data using the following steps:
    - Scaling was applied to the continuous variables of the test data.
    - Utilizing the built model and a cutoff of 0.2, predictions were generated for the test dataset.
    - A confusion matrix was created, resulting in an accuracy of 92%, sensitivity of 88%, and specificity of 94%.
    - These results indicate that the model performs well on unseen data.
    - Lead conversion scores were assigned to each lead by multiplying the conversion probability by 100.
    - The most significant features influencing the conversion probability were identified.

Key learnings from this assignment include:

- Understanding the process of data exploration and handling missing values.
- Recognizing the importance of performing EDA and data pre-processing.
- Implementing a systematic approach for model building and feature selection, considering the impact on both training and test datasets.

Successfully solving problems through teamwork and leveraging individual strengths.