Summary Report

In order to address the problem of selecting the most promising leads for X Education Company, we applied Logistic Regression to the given dataset. Our approach consisted of the following steps:

1) Data loading and preparation: We loaded the data into a Jupyter notebook and performed missing value and outlier treatment for all columns. We also dropped columns that had more than 45% missing values.

2) Exploratory Data Analysis (EDA): We conducted EDA on each categorical column to visualize data imbalance and dropped high data imbalanced columns.

3) Train-test split and feature scaling: We performed train-test split using the sklearn library and scaled the train data using Standard Scaler.

4) Feature selection: We used Recursive Feature Elimination (RFE) to select 20 features for the model.

5) Logistic Regression: We used the statsmodels GLM method to perform Logistic Regression on the selected features and checked coefficients, p-values, and Variance Inflation Factor (VIF).

6) Model evaluation: We evaluated the model using accuracy, specificity, and sensitivity. We also plotted the Receiver Operating Characteristic (ROC) curve to check the balance between True Positive Rate (TPR) and False Positive Rate (FPR).

7) Cut-off selection: We selected the optimal cut-off point (0.34) for lead conversion by plotting confusion matrix for different probability cut-offs.

8) Re-evaluation: We re-evaluated the model using the selected cut-off point and found an accuracy of 90.6%, precision of 86.4%, and recall of 90%.

9) Model testing: We scaled the test data and performed prediction of lead conversion. We evaluated the prediction on test data by accuracy, specificity and sensitivity matrix, and we found accuracy -91.1%, precision – 85.6% and recall – 91.9%.

10) Lead Score Calculation: We created lead conversion score = (conversion probability * 100) to give a score between 0 to 100 where higher the value means the lead is “hot” and there is high possibility that the lead can be converted.

Throughout this assignment, we learned various techniques for handling missing values, outliers, and data imbalance, as well as how to use Python libraries for logistic regression and feature selection. Additionally, we gained experience in balancing sensitivity and specificity for model selection and team collaboration for problem-solving.