SUMMARY REPORT

To solve the lead conversion efficiency problem for X Education, we developed a logistic regression model using Python and applied it to identify "Hot Leads" — leads with a high probability of conversion.

Step-by-Step Process:

1. **Data Import & Exploration**
    - Loaded the leads dataset in Jupyter Notebook.
    - Understood variable types and assessed the distribution of target variable Converted.
2. **Missing Value & Outlier Handling**
    - Dropped columns with over 70% missing values.
    - Replaced categorical 'Select' values with NaN.
    - Removed irrelevant and low-variance features.
    - Treated outliers in numeric variables where necessary.
3. **EDA (Exploratory Data Analysis)**
    - Performed univariate and bivariate analysis.
    - Dropped highly imbalanced categorical variables (e.g., >95% one class).
4. **Data Preparation**
    - Applied get_dummies() for categorical variables.
    - Scaled numeric features using StandardScaler.
    - Split the dataset into training and testing sets (70:30 ratio).
5. **Feature Selection & Model Building**
    - Applied Recursive Feature Elimination (RFE) to select top 15 features.
    - Built a logistic regression model using statsmodels.GLM().
    - Iteratively removed features with high p-values and multicollinearity (checked using VIF).
6. **Model Evaluation – Train Set**
    - Used 0.5 as initial cutoff; evaluated performance using confusion matrix.
    - Metrics at 0.5 cutoff:
        - Accuracy: 92%
        - Sensitivity: 86%
        - Specificity: 96%
    - Plotted ROC Curve; AUC: **0.96** (excellent model quality).

7. **Cutoff Optimization**
    - Analyzed performance across thresholds from 0.0 to 0.9.
    - Chose 0.2 as the optimal cutoff for balanced sensitivity and specificity.
8. **Model Validation – Test Set**
    - Transformed test data with the same scaler.
    - Evaluated model using the optimized cutoff (0.2).
    - Test set metrics:
        - Accuracy: 92%
        - Sensitivity: 87%
        - Specificity: 94%
    - Model performed consistently on unseen data.
9. **Lead Scoring**
    - Final lead score = predicted probability × 100
    - Helps business rank and target leads more effectively.

Key Learnings

- Effective handling of missing data and outliers.
- Importance of categorical encoding and scaling.
- Use of logistic regression for classification.
- Model evaluation using ROC, sensitivity, and specificity.
- Business insight generation through lead scoring.
- Collaboration and end-to-end problem solving in a real-world context.