Problem Statement

Company Context: X Education, a leader in providing online courses to industry professionals, draws numerous leads through its website, marketing campaigns, and referrals. Despite this, the current lead conversion rate is only about 30%, which is suboptimal for the company's growth ambitions.

Challenge: To significantly boost their conversion rate to the target of 80%, X Education needs a sophisticated solution that accurately identifies and prioritizes high-potential leads, thereby optimizing the follow-up process and reducing efforts on low-probability prospects.

Objective: Develop an advanced Machine Learning model capable of predicting the likelihood of each lead converting into a customer. This model will enable the sales team to focus their efforts strategically on leads with the highest potential, enhancing efficiency and driving up the conversion rate.

Expected Impact:

- Achieve a target conversion rate of 80%.
- Streamline the sales process by focusing on high-potential leads.
- Optimize resource allocation, saving time and increasing sales effectiveness.
- Drive greater sales and revenue growth for X Education.

Workflow:

1. **Data Loading and Cleaning**:
    - **Data Acquisition**: Imported a dataset with 37 diverse features and 9,240 records.
    - **Value Replacement**: Substituted 'select' entries with NULL to denote unselected options.
    - **Missing Value Handling**: Applied a 45% threshold to discard features with excessive missing values; imputed remaining gaps using appropriate aggregate functions post-analysis.
    - **Outlier Management**: Identified and mitigated outliers using capping techniques.
    - **Feature Elimination**: Removed non-essential columns following detailed examination.
    - **Data Correction**: Rectified incorrect entries and consolidated low-frequency values for simplified analysis.

2. **Exploratory Data Analysis**:
    - **Conversion Rate Check**: Lead conversion rate identified as 38%.
    - **Univariate Analysis**: Conducted on both numeric and categorical features to extract valuable insights.
    - **Correlation Insights**: Used a heatmap to visualize correlations among numeric features.

3. **Preparing the Data for Modelling**:
    - **Binary Mapping**: Converted binary values to 0 and 1.
    - **Dummy Variables**: Created for categorical features.
    - **Train-Test Split**: Split data into 70% training and 30% testing sets.
    - **Rescaling**: Applied fit-transform to rescale continuous variables in the training set.

4. **Training and Modelling the Data**:
    - **Correlation Analysis**: Identified top correlation pairs among independent variables.
    - **Feature Selection**: Used Recursive Feature Elimination (RFE) for feature ranking and selection.
    - **Manual Elimination**: Removed variables with high p-values (>0.05) and high VIF (>5).
    - **Model Selection**: Chose the final model based on key statistics, significant variables, and absence of multi-collinearity.

5. **Prediction and Model Evaluation**:
    - **Train Set**:
        - **Initial Predictions**: Made predictions with the final model, set random cut-off, and evaluated using metrics (Accuracy, Sensitivity, Specificity, Precision, Recall).

        - **Optimal Cut-Off**: Determined as 0.3, re-evaluated metrics, and assigned lead scores.
    - **Test Set**:
        - **Transformation**: Applied Standard Scaler to test data's numeric features.
        - **Alignment**: Ensured X test features aligned with X train.
        - **Final Predictions**: Made predictions with the final model using 0.3 cut-off, evaluated using various metrics.

Evaluation Score Chart

- **Learnings**:
    - Evaluation metrics for both train and test sets show strong performance, ensuring model stability and suitability for achieving business goals.