Introduction:

In competitive online education sector, maximizing lead conversion is crucial for business growth. X Education, which sells online courses wants to improve its leads conversion rate from 30% to 80%. This assignment employs data science to uncover influential factors that enhance lead-to-enrolment conversion, enabling data-driven marketing strategies. Below mentioned are steps taken and learnings from the assignment.

1. Data Analysis

The dataset comprised of 9,200 records with multiple categorical, few numerical variables and target variable - Converted. The data analysis phase involved:

- *Variable Exploration:* Examined the distribution and data types to understand the dataset's composition.
- *High-Cardinality Features:* Identification of features with numerous unique values. Consolidation of categories improved model interpretability
- *Placeholder values*: "Select" were marked as missing
- *Variance analysis*: To eliminate low-variance features

2. Data Cleaning and Transformation

The dataset underwent several cleaning and transformation steps:

- *Binary Encoding*: Categorical variables with two levels were converted to binary (0 and 1)
- *Feature Reduction*: Features with low variance (Country,Magazine,NewspaperWWArticle) were removed
- *Category Consolidation*: High-cardinality categorical features (Specialization) were merged for simplicity
- *Outlier Treatment*:
    - IQR Method: Applied to skewed distributions (selected)

3. Exploratory Data Analysis (EDA)

EDA provided valuable insights into the data distribution and their impact on lead conversion:

- *Proportion Analysis*: Mapping category proportions highlighted significant contributors such as 'References', 'Welingak Website'
- *Behavioural Insights* such as:
    - Features like 'Time Spent on Website' positively correlated with conversions, emphasizing their importance as potential drivers.
    - Working professionals are the best occupation to target

5. Feature Engineering and Scaling

- *Dummy Variable Creation*: Dummy variables were generated for categorical features with more than two values
- *Scaling*: Numerical variables were normalized using Standard Scaler to ensure uniformity during model training.

6. Model Development and Evaluation:

A logistic regression model was developed using the following techniques:

- *Feature Selection*:
    - Recursive Feature Elimination (RFE) identified the top 15 variables.

    - Further refinement was done using Variance Inflation Factor (VIF < 5) and p-values (< 0.05).
- *Performance Metrics*:
    - Confusion matrix analysis showed accuracy and sensitivity of around 80%.
    - ROC curve analysis identified an optimal cutoff of 0.375, prioritizing sensitivity (recall) due to the higher cost of missed leads.

7. Key Findings / Insights:

The model revealed critical factors influencing lead conversion:

- *Critical Factors*:
    - Tags provide good information - Closed by Horizzon, Lost to EINS, Will revert after reading the email have high positive impact
    - Leads who are in touch - "Last Notable Activity_Had a Phone Conversation", :Last Notable Activity_SMS Sent" - convert more. Customer Interaction matters.
- *Professional Background*: Working professionals were more likely to enrol, highlighting a key target segment

8. Conclusion/ Recommendations:

This assignment identifies strategic factors influencing lead conversion for X Education, enabling targeted marketing interventions. Recommendations include:

- **Targeted Outreach:** Take the inputs from Logistic regression to target leads identified by high positive impact factors. Customize communication strategies for high conversion rate segments.
- **Enhanced Engagement Tactics**: Personalizing website experiences to increase user engagement
- **Optimized Marketing Channels**: Focusing on high-converting sources like References, Welingak Website'