Summary Report

Problem Statement:

- X Education, an online education platform, is facing a low lead conversion rate of **30%**.
- The goal is to develop a **lead scoring model** that identifies "Hot Leads"—leads with a higher probability of conversion.

Business Objectives:

- Build a logistic regression model to score leads by conversion likelihood.
- Increase lead conversion rate to 80%.
- Ensure the model adapts to future business changes.

Steps Taken:

1. Data Reading & Understanding:

- The dataset contains 9,240 leads and 37 columns, with "Converted" as the target variable indicating customer conversion (1) or not (0).

2. Data Cleaning:

- **Duplicate Records:** No duplicates found, ensuring data integrity.
- **Missing Values:**
  - 'Select' values were replaced with NaN.
  - Columns with over 35% missing data were dropped.
  - Remaining missing values were imputed using the mode or Unknown.

3. EDA (Data Visualization):

- **Univariate Analysis:** The conversion rate was 38.54%, indicating room for improvement.
- **Bivariate & Multivariate Analysis:** Relationships with the target were analyzed, noting that leads with “SMS Sent” as the last activity had higher conversion rates.
- **Outlier Treatment:**
  - Outliers in 'TotalVisits' and 'Page Views Per Visit' were identified using boxplots.

- Capped outliers at the 95th percentile to minimize their impact.

4. Data Preparation:

- Created dummy variables for categorical features.
- Split the data into training and test sets.
- Applied feature scaling to ensure all features were on the same scale.

5. Model Building:

- **Feature Selection:**
  - Applied Recursive Feature Elimination (RFE) to select 15 key features.
  - Dropped columns with p-values > 0.05.
  - Ensured no multicollinearity using VIF, retaining 13 features Applied RFE to select 15 key features.

6. Model Evaluation:

- **ROC Curve & AUC:**
  - AUC of 0.89, indicating strong performance in distinguishing positive and negative .
  - Using an optimal cutoff of 0.34, the model achieved approximately 80% accuracy, sensitivity, and specificity on both train and test data.
- **Precision Recall Trade-off:** It suggests 0.41 as the threshold, but sensitivity drops ,proceed with the 0.34 threshold .

Conclusion:

Key Features Driving Conversion:

- **Total Time Spent on Website (4.4841):** Higher engagement on the website leads to better conversion rates.
- **Lead Origin - Lead Add Form (3.7886):** Form submissions are more effective in generating high-converting leads.
- **Occupation - Working Professional (3.6383):** Campaigns targeted at professionals are more likely to convert.

Additional Insights:

- Phone Conversations and Olark Chat are strong predictors of conversion.
- Leads from Welingak Website show higher conversion potential.
- SMS Follow-ups improve the likelihood of conversion.

Recommendations for X Education:

- **Enhance Engagement:** Increase website interaction with personalized content.
- **Optimize Lead Forms:** Improve and streamline forms for higher-quality leads.
- **Target Professionals:** Focus campaigns on working professionals.
- **Leverage Communication:** Use phone calls and live chat to increase conversion rates.
- **Utilize SMS Follow-ups:** Increase conversion rates with SMS follow-ups.

Implementing these strategies will help X Education increase lead conversion rates, enhance sales efficiency.