Analysis Summary for X Education

This analysis was conducted to help **X Education** attract more industry professionals to enrol in their courses. By leveraging data on customer interactions, we aimed to uncover key factors influencing course enrolments and optimize marketing efforts.

Key Findings & Approach

1. Data Cleaning

- Many of the categorical variables have a level called 'Select' which needs to be handled because it is as good as a null value, as these values are not provided by customers
- Columns with unique values equal to one were dropped, as there is nothing to compare with and it won't affect our analysis
- Columns with missing values more than 40% were dropped
- For some columns with high missing values, a new level "Not provided was created as by dropping these columns we would lost lot of information crucial for our analysis
- Highly skewed categorical columns with poor conversion rates were dropped

2. Exploratory Data Analysis (EDA)

EDA revealed that data for 2 numerical column were highly skewed, we cap the outliers to 95%. The missing values were imputed with median to preserve the original distribution. Many categorical variables contained irrelevant or redundant elements. These were refined to ensure meaningful insights during modelling.

3. Feature Engineering

- **Binary Mapping**
- New levels were created for some features, some levels were merged and also a new column(Tag_category with levels High, Medium, low) was created to reduce dimensionality
- **Dummy Variables:** Categorical variables were transformed using one-hot encoding.
- **Scaling:** Numeric values were standardized using **StandardScaler**

4. Data Splitting

The dataset was split into **70% training** and **30% testing** to ensure robust model validation.

5. Model Building & Optimization

- **Feature Selection:** RFE was applied to identify the **top 20 relevant variables**. Further refinement was done by manually eliminating variables based on **Variance Inflation Factor (VIF < 5)** and **p-values (< 0.05)** to reduce multicollinearity and improve model efficiency.
- **Logistic Regression Model:** Built and refined using statistically significant features.

6. Model Evaluation

- **Precision-Recall Trade-off:** Adjusting the cut-off to **0.41** optimized **precision (90%)** and **recall (82%)**, balancing false positives and false negatives.

- **High AUC (0.96-0.97)** indicates the model is highly effective at distinguishing between potential and non-converting leads.
- **Confusion Matrix & ROC Curve:** The **optimal cut-off threshold (0.3)** was identified, leading to an accuracy, sensitivity, and specificity of approximately **90%** on the training data.
- **Test Set Prediction:** The model achieved **90% accuracy, sensitivity, and specificity** on unseen test data.

Top Recommendations for X Education

- **Focus on leads from the Welingak Website and SMS Sent category.**
- **Prioritize working professionals**, as they show high conversion intent.
- **Increase conversion rates for leads from API & Landing Page Submissions.**
- **Use CRM to automate follow-ups and optimize call strategies** based on conversion probability.
- **Adjust model threshold based on business goals:**
    - If maximizing conversions → **Lower threshold (0.3)** to improve recall.
    - If minimizing unnecessary calls → **Increase threshold (0.7-0.8)** to improve precision.

By implementing these **data-driven recommendations**, X Education can **improve lead conversion rates, optimize outreach efforts, and maximize enrolment success**.