This analysis was conducted for **X Education** to identify strategies to attract more industry professionals to their courses. The dataset provided detailed information on potential customer behavior, including how they reached the site, the time spent on the site, and conversion rates. Below are the steps taken during the analysis:

1. Data Cleaning:

- The dataset was mostly clean, with a few null values. We replaced the "Select" option with nulls as it provided no meaningful information.
- Columns with over 40% null values were dropped, and values were imputed for columns with less than 40% nulls.
- Some columns were removed due to data imbalance.

2. Exploratory Data Analysis (EDA):

- We performed univariate and bivariate analyses on both continuous and categorical variables.
- It was observed that the distribution of data in numerical columns was not normal, and outliers were treated accordingly.

3. Dummy Variables:

- Dummy variables were created for categorical columns.
- For numerical features, we applied **StandardScaler** to standardize the data.

4. Train-Test Split:

- The dataset was split into 70% training data and 30% test data.

5. Model Building:

- A **logistic regression** model was used for this classification problem.
- **Recursive Feature Elimination (RFE)** was employed to select the most relevant variables. Variables with **VIF ≤ 5** and **p-value ≤ 0.05** were retained, while others were removed.
- An **ROC Curve** was generated to illustrate the tradeoff between sensitivity and specificity, and a line graph was used to identify the optimal cutoff probability for balancing these metrics.
- We also analyzed the tradeoff between **precision** and **recall**.
- Predictions were made on the test data, and confusion matrices were created for both the train and test datasets.
- Key metrics such as **Accuracy, Sensitivity, Specificity, Precision, Recall**, and the **F1 Score** were calculated.

6. Conclusion:

- **Time on site**: Leads spending more than average time on the site are promising prospects, suggesting that targeting them could increase conversions.
- **SMS Marketing**: SMS messages have a high impact on lead conversion.

- **Landing Pages**: Submissions through landing pages are an effective way to capture leads.
- **Specializations**: Leads from marketing management and human resources management specializations show higher conversion rates.
- **Referrals**: Offering incentives for referrals can generate more leads.
- **Alerts**: Sending alert messages or notifications has a significant positive effect on lead conversion.
- **Threshold Selection**: The optimal threshold was determined based on a balance of Accuracy, Sensitivity, Specificity, Precision, and Recall.

Model Performance:

- The model achieved an **88% accuracy** on both train and test datasets.
- It demonstrated a **Recall rate of 83%** and **Precision of 87%**, indicating that the model is effective at identifying high-potential leads (Hot Leads).
- Overall, the **Logistic Regression model** proves to be a good fit, with strong performance metrics and practical insights for improving lead conversions.