EXECUTIVE SUMMARY – LEAD SCORING CASE STUDY

The aim of this study is to build a model for company X Education to convert potential users and increase the conversion rate. The data was validated to target the correct group using the following steps:

1. **EDA**:
    - Checked for null values and dropped columns with more than 45% missing values.
    - Replaced NaN values with 'not provided'.
    - Imputed all 'not provided' values with India since it was the most common occurrence.
    - Dropped the 'India' column.
    - Worked on numerical variables, outliers, and dummy variables.
2. **Train-Test Split & Scaling**:
    - The data was split into 70% train and 30% test data.
    - Min-max scaling was applied to the variables ['TotalVisits', 'Page Views Per Visit', 'Total Time Spent on Website'].
3. **Model Building:**
    - RFE was used for feature selection.
    - The top 15 relevant variables were selected using RFE.
    - The remaining variables were removed manually based on VIF values and p-values.
    - A confusion matrix was created, and the overall accuracy was 80.91%.
    4. Model Evaluation:

- **Sensitivity-Specificity:**
    - On training data:
        - The optimum cut-off value was found using the ROC curve, and the area under ROC curve was 0.88.
        - The optimal cut-off value was found to be 0.40, which gave an accuracy of 86.91%, sensitivity of 85.94%, and specificity of 87.50%.
    - On test data:
        - The model yielded an accuracy of 87.02%, sensitivity of 85.23%, and specificity of 86.50%.
- **Precision-Recall**:

- On training data:
    - The optimal cut-off value of 0.40 was found, which yielded an accuracy of 86.80%, precision of 81.71%, and recall of 85.32%.
- On test data:
    - The model yielded an accuracy of 87.57%, precision of 82.87%, and recall of 85.26%.

5. Based on the evaluation, if the Sensitivity-Specificity evaluation is preferred, the optimal cut-off value would be 0.40.

CONCLUSION

The following variables are the top contributors to conversion, as identified by the model:

- Lead Source: Total Visits, Total Time Spent on Website
- Lead Origin: Lead Add Form
- Lead Source: Direct Traffic, Google, Welingak Website, Organic Search, Referral Sites
- Last Activity: Do Not Email_Yes, Last Activity_Email Bounced, Olark Chat Conversation

The model performs well in predicting conversion rates and provides the company with the confidence to make informed decisions based on its predictions.

RECOMMENDATIONS:

- Call leads from lead sources Welingak Websites and Reference
- Call working professional leads
- Call leads who spent more time on the websites
- Call leads from lead source Olark Chat
- Do not call leads whose last activity was Olark Chat Conversation
- Do not call leads whose lead origin is Landing Page Submission
- Do not call leads whose specialization was Others
- Do not call leads who chose the option of Do not Email as Yes