Summary Report: Lead Scoring Case Study

Approach and Learnings

Approach Used:

1. **Data Loading and Initial Analysis:**
    - The dataset, containing 9240 entries and 37 columns, was loaded using pandas.
    - It includes features like Prospect ID, Lead Number, Lead Origin, Lead Source, etc.
2. **Data Cleaning:**
    - Duplicates were checked (Lead Number and Prospect ID) but none were found.
    - Columns with only one unique value (e.g., Magazine) were removed.
    - Missing values treatment: Replacing 'Select' with NaN, NaN converted to 'not provided' for specific columns.
    - Columns with > 45% missing data dropped, the country column was bucketed into “India” and “Outside India”.
3. **Exploratory Data Analysis (EDA):**
    - Univariate analysis performed using count plots for categorical variables and checking for class imbalance.
    - Bivariate analysis done between categorical variables and the target variable 'Converted'.
    - Multivariate analysis examined numerical variables through frequency distributions and correlation analysis.
    - Outliers were detected and removed using boxplots and the IQR range method.
4. **Feature Engineering:**
    - Dummy variables were created for categorical variables, and numeric features were scaled using Standard Scaler.
    - The dataset was split into training and test sets.
5. **Model Building:**
    - Recursive Feature Elimination (RFE) was used for feature selection.
    - Logistic Regression model applied, a Generalized Linear Model (GLM) with Binomial family (binary classification).
    - Multiple iterations refined the model by removing features with high p-values and high Variance Inflation Factor (VIF).

6. **Model Evaluation:**

- The final model was applied to the training set, predicting probabilities that were converted into Lead Scores by multiplying by 100.
- Model performance was evaluated using confusion matrix: accuracy, sensitivity, specificity, precision, and recall at a cutoff of 0.5.
- The Receiver Operating Characteristic (ROC) curve was plotted, with a ROC value of 0.96 for model goodness.
- Accuracy, Sensitivity, and Specificity were plotted for various thresholds. This and the Precision-Recall curve helped determine the optimum cutoff of 0.3.
- The model, using the cutoff of 0.3, was run on the test set and evaluated with the same metrics. The results included Lead Number to associate predictions with exact leads.

Insights:

- **Data:** Most leads originated from India and came via landing page submissions from Google. Many leads had no city information, though Mumbai was the second most frequent city. Most leads were fresh (not referred), unemployed, and sought career enhancement through courses. Conversions were higher via SMS. "Do Not Disturb" preferences were less likely to convert. Leads spending more time on the website were more likely to convert, and visits and page views per visit were highly correlated.
- **Model:** 'Tags' emerged as a key feature, with top positive coefficients for "Tags_Closed by Horizzon," "Tags_Lost to EINS," and "Tags_Will revert after reading the email." The lead source "Welingak Website" and "SMS sent" as last activity were also important predictors.
- **Model Performance:** The final model had an accuracy of approximately 92%, with **sensitivity/recall at 89%** and specificity at 94%. The model performed well on both training and test sets, making it effective for lead scoring.