Lead Scoring Analysis - Summary Report

1. Data Preparation and Cleaning

The analysis begins by importing the dataset and performing an initial inspection to understand its structure. This includes identifying missing values, handling null entries, and cleaning categorical and numerical variables. Unnecessary columns are dropped, and categorical variables are encoded as dummy variables for modeling.

2. Exploratory Data Analysis (EDA)

EDA is conducted to analyze trends and relationships within the data.

- **Univariate Analysis**: Both categorical and numerical variables are assessed individually.
- **Bivariate Analysis**: The relationship between different categorical variables and conversion rates is explored.
- Outliers are examined but found to be negligible, so no extensive outlier treatment is performed.

3. Feature Engineering

- Dummy variables are created for categorical variables to ensure compatibility with machine learning models.
- Features with high multicollinearity are identified using **Variance Inflation Factor (VIF)** and **Recursive Feature Elimination (RFE)**.
- Only statistically significant variables are retained in the model.

4. Model Building

- The dataset is split into training and testing sets to ensure robust model performance.
- Logistic Regression is used as the primary classification model.
- The model undergoes iterative refinement by removing variables with high p-values (>0.05) and ensuring VIF values are within an acceptable range.

5. Model Evaluation

- The model achieves an accuracy of **81%**, which indicates good performance.
- **Sensitivity (Recall):** ~70% (Ability to correctly identify positive leads)
- **Specificity:** ~87% (Ability to correctly reject non-converting leads)

6. Optimization via ROC Curve

- The default cutoff of 0.5 is adjusted using the **ROC Curve** to find an optimal threshold for better sensitivity and specificity balance.

- This helps improve lead conversion predictions by adjusting false positives and false negatives based on business requirements.

7. Conclusion and Business Impact

- The model provides an effective lead-scoring mechanism, helping prioritize high-potential leads.
- With optimized sensitivity and specificity, it ensures that sales teams focus on the right leads, improving efficiency and revenue.
- Further improvements can be made by incorporating additional business-specific features or trying advanced classification models.

This structured approach ensures the lead-scoring model is both **accurate and actionable**, enhancing decision-making in lead prioritization.